

No new start at Los Alamos

A fresh contract for the management of the New Mexico nuclear-weapons laboratory offers it little prospect of a happy and prosperous new year.

Just before Christmas, an epic struggle for management control of the United States' most venerable nuclear-weapons laboratory came to a close. The result, rather to the surprise of many observers, was that the University of California will retain the management contract to run the Los Alamos laboratory (see page 8), in partnership with the engineering group Bechtel and two other private corporations.

This means that control of the New Mexico laboratory will edge towards the private sector. But the more radical option of passing it on to fresh management under a rival consortium led by the University of Texas has been rejected. Once, this news would have led to celebrations among the 8,000 or so University of California staff at Los Alamos. But their mood is instead forlorn. Staff pensions and other benefits are not guaranteed under the new arrangement, and recent actions by the University of California have eroded goodwill.

The process by which the Department of Energy awarded the contract has been murky, even by the usual standards of such exercises. Few believe that the department's grey-suited administrators really made an independent choice. Rather, the process was characterized by delays and heavyweight political lobbying from Senator Pete Domenici (Republican, New Mexico), among others. That's par for the course, as the 'management crisis' at Los Alamos has always been more about Washington politics than about actual administrative issues at the lab.

Ever since Wen Ho Lee was accused of espionage at the lab in 1999 (he was later acquitted of major charges and convicted of minor regulatory infringements), a group in Congress led by Joe Barton (Republican, Texas) has relentlessly sought to impugn the laboratory's staff and its management by the University of California. The campaign echoes previous efforts to bring Los Alamos scientists under tighter administrative, and perhaps military, control. Its proponents have overplayed security issues at the laboratory and implied that senior scientists there cannot be trusted, either in administration or security. They have issued a stream of overblown

rhetoric, leading to the brief and unfortunate appointment of Pete Nanos, a former naval officer, to run the laboratory, as well as to last year's tendering process.

The University of California has been contracted to run the laboratory since 1943, and it traditionally did so for a nominal fee, kept its hands off day-to-day management, and offered scientists and engineers there the opportunities that came with affiliation to one of the world's best public university systems.

By seeking to blame the university for the lab's difficulties, some in Washington have sought to deflect attention from their own culpability, which is considerable. In reality, the laboratory is controlled not by its contractor, but by the byzantine Department of Energy and its overseers in several congressional committees. Because these committees are happy to make rules but are incapable of constraining expenditure, the nuclear-weapons labs have lately been given more money and much more oversight and regulation. As a result, they became steadily less efficient and productive.

The University of California and its partners, meanwhile, must pick up the pieces at Los Alamos and start afresh. Their appointment of Michael Anastasio, director of the rival Lawrence Livermore laboratory, to run Los Alamos has not exactly thrilled the existing staff there, given the historical rivalry between the two institutions. And staff briefings just before Christmas shed little light on what the new management team is actually going to do.

Los Alamos retains expertise in areas such as physics, materials, computer science, neutron scattering and mathematics. The key to its continued relevance is close liaison between its researchers in these fields and the academic community outside. But given the constraints under which they must operate, the new contractors will be hard-pressed to make the laboratory thrive. ■

"By seeking to blame the university for the lab's difficulties, some in Washington have sought to deflect attention from their own culpability."

Developing resistance

A study of opposition to a vaccine for children shows how the public can lose faith in science.

Nostradamus we are not, but a safe prediction for 2006 is that initiatives promoting public engagement in science and technology policy-making will proliferate. There will, of course, be devils in the details, and critical assessments will be required. But *Nature*, having consistently championed public engagement, can nevertheless only welcome its development.

But there are times when no amount of explanation and consultation can counter the resistance of some sectors of the public, often representing a strong current in society, to the most carefully crafted science-based advice. Because the stakes for people's quality of life, economic development and the rights of individuals can be high, governments and the rest of us need to understand how and why such resistance to science develops.

Studies by social scientists have a major role to play in providing an understanding of how such resistance develops. A notable example is British research led by Melissa Leach at the University of Sussex into strong resistance by parents to their children receiving a freely available vaccination against measles, mumps and rubella



(MMR). The relevance of this research extends well beyond the particular circumstances and region studied.

The context was an outbreak in 1998 of public concern in Britain that the MMR vaccination might lead to autism in young children. The worries, which were stimulated by some scientists, proliferated and persisted despite increasingly robust reassurances from the government and clinicians that there was no epidemiological evidence for such a link.

The study of parents' responses provides a vivid demonstration of how people outside the relevant research communities develop their own knowledge and interpretation of the literature, and how the Internet allows this 'lay expertise' to be shared within a climate of shared perceptions of risk (see the unrefereed report *MMR Mobilisation: Citizens and Science in a British Vaccine Controversy*; www.ids.ac.uk/ids/bookshop/wp/wp247.pdf). For example, one parent, David Thrower, compiled his own review of the literature (www.whale.to/a/thrower04.html), and several websites promptly supplied hostile responses to every government reassurance.

Leach's research highlights the influences on such resistance. Confounding stereotypes, the parents ranged across the social classes and in many cases displayed a sophisticated understanding of the issues. Surveys of parents on both sides of the debate revealed a significant association of those opposed to MMR with family histories of illness and with an interest in alternative medicines and homeopathy. There was a strong sense of responsibility among mothers of both persuasions about decisions over whether to have their children vaccinated, with concerns about the social value of vaccination in terms of herd immunity being given much lower priority.

The study's account of these developments is only partial, however. Close reading reveals an undercurrent of sympathy towards the parents and relatively cursory attention to attempts by the govern-

ment and others to survey the evidence and respond to the parents' concerns. But the research is focused on the parents by definition, and provides an important starting point in trying to understand the various forces at work in a classic stand-off between citizens and science-based government advice.

Leach's work points to a conflict between concerns about MMR based on individual clinical studies versus government reassurances based on epidemiology. Soon after the publication of Leach's report, a meta-analysis of the literature on MMR by the prestigious Cochrane Collaboration, while highlighting shortcomings in many studies, concluded that there was a lack of evidence to support a link with autism (www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD004407/pdf_fs.html).

A visit to one of the websites opposed to the MMR vaccine (www.jabs.org.uk) reveals a critique of the meta-analysis that attempts to undermine its reassurance. And so the debate continues. Meanwhile, the uptake of MMR vaccine, which fell significantly, is recovering.

In most countries, the departments of health are responsible for advising citizens on health matters. They, perhaps, have the most to gain from in-depth studies of public responses, while scientific academies may also find such accounts of alternative science sobering. But as happened with MMR, some key sectors of the public can lose faith in both the government and scientists. Thus there is a strong case for a well-resourced independent national agency that commands the trust of both the government and the public in matters of health protection and is empowered to take responsibility for mediating in such debates. ■

"There is a strong case for a well-resourced independent agency that commands the trust of both the government and the public."

Sound science

Audio files downloaded from the Internet can enrich scientific communication.

Thanks in large part to the ubiquity of Apple's iPod, the Internet is now host to a new kind of audio transmission — the podcast.

Last September, we quietly introduced the *Nature* Podcast, which each week highlights a selection of papers and news features from the latest issue, with interviews of authors and their peers. In this way we let scientists explain their results to a wide audience, in their own words. Their input is augmented by comment and analysis from our own editors and journalists.

For the uninitiated, a podcast — a nifty contraction of iPod and broadcast — involves the automatic downloading of an audio show via an Internet content-distribution mechanism known as an RSS feed. Listeners just enter the address of a show in the podcast section of Apple's iTunes software. Each time a show is released, it is downloaded straight to their computer.

The podcast has come to the fore because of the iPod's success and the convenience of hearing an audio item at a time that suits the listener, rather than when radio schedules dictate. Podcast fans can

now catch up on *Nature* when setting up their experiments, sitting in traffic or walking the dog. They can also enjoy a growing array of other content, ranging from podcast museum tours to directors' commentaries that augment television programmes.

Nature is immensely pleased that, as we go to press, our show sits unassumingly between a Bob Dylan commentary and the CNN news update in iTunes' top 100 podcast chart. This demonstrates how the technology is helping the work that we publish to reach a wider public. There are other science podcasts too, including contributions from the *New England Journal of Medicine* and from NASA, as well as podcast versions of established radio shows, such as *Science Friday*. The presence of science podcasts in the charts suggests that there is plenty of interest in their subject matter.

After experimenting successfully with the *Nature* Podcast over the past few months, we are now establishing it as a more permanent component of our publication. We warmly invite readers to listen to the show now at www.nature.com/nature/podcast and then take part in our listener survey to let us know how the new venture can be further enhanced. ■

"The presence of science podcasts in the charts suggests that there is plenty of interest in their subject matter."

RESEARCH HIGHLIGHTS

P. MORRIS/ARDEA.COM



Sharp-eyed observation

J. Exp. Biol. **209**, 18–25 (2005)

The slit-shaped pupils of animals from cats to geckos (pictured) evolved to complement the optical properties of their lenses, say researchers.

Ronald Kröger and his student Tim Malmström, of Lund University in Sweden, examined the eyes of 20 vertebrates with an infrared camera. They found slit pupils only in animals with multifocal lenses. These focus light of different wavelengths through different concentric zones, producing a sharper image than a lens with a single focal point at the centre, such as those of humans.

In a multifocal lens, a circular pupil would contract to obscure entire concentric regions needed to focus some wavelengths, whereas with a slit-shaped pupil, light always passes through a portion of each concentric ring.

CELL BIOLOGY

Bound to change

Science **310**, 1966–1970 (2005)

Vioxx, the painkiller recently withdrawn amid concerns about its side effects, works by inhibiting the enzyme cyclooxygenase-2, or COX-2, to limit inflammation. Now researchers in the United States have identified an enzyme in a separate inflammatory pathway that activates COX-2.

Solomon Snyder of the Johns Hopkins University School of Medicine in Baltimore, Maryland, and his colleagues found in immune cells from mice that inducible nitric oxide synthase specifically binds to and activates COX-2. Next, they showed that small peptides can disrupt this interaction. The group suggests that drugs designed to interfere with this pathway in humans could offer a new anti-inflammatory treatment.

CHEMISTRY

Bright prospects

Nature Meth. **3**, 35–40 (2005)

A design for a molecular cage that can be unlatched by light offers promise for *in vivo* experiments, say its inventors.

Light-sensitive molecules called chromophores have been used for decades to cage small compounds, because the release of their contents can be precisely controlled. The cages disgorge their cargo when pulsed with light at critical moments in a synthesis reaction, for example, or inside living cells.

Graham Ellis-Davies of Drexel University College of Medicine in Philadelphia, Pennsylvania, and his colleagues have come

up with nitrodibenzofuran, which opens in response to light up to 160 times more efficiently than existing chromophores. This offers advantages for experiments in cells, as these can be damaged by strong doses of light. The researchers used the chromophore to release calcium ions in heart cells.

STEM CELLS

Plant matters

Cell **123**, 1337–1349 (2005)

All multicellular organisms require a supply of stem cells that can differentiate into any type of cell that needs to be replaced or supplemented in a lifetime.

Three different retinoblastoma genes have been implicated in the maintenance of mammalian stem cells, but it has not yet been possible to determine whether they influence stem cells directly, or their daughter cells.

So Ben Scheres from Utrecht University in the Netherlands and his colleagues turned to

plants, in which stem cells are easier to see. In the root pictured below left, the stem cells are located just below the green dots. The team shows that the retinoblastoma pathway plays a similar role in maintaining 'stem cellness' in plants, and that its influence is directly on stem cells, and not their progeny.

CELL BIOLOGY

Quick fix

Mol. Cell **20**, 783–792 and 793–799 (2005)

DNA is easily damaged, so the cell has evolved numerous DNA-repair mechanisms to avert disaster. Some types of damage block normal DNA replication, but can be fixed by a process called homologous recombination repair. The question is: which of the 20 or so different polymerase enzymes in the cell is involved in this process?

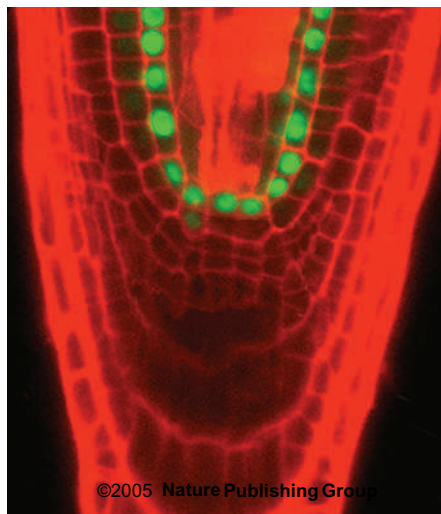
Using different approaches, two groups have independently solved the mystery. Stephen West of Cancer Research UK in London and his colleagues used biochemical approaches to identify the enzyme as DNA polymerase η , whereas Shunichi Takeda of Kyoto University in Japan and his team used genetics to reach the same conclusion.

ORGANIC CHEMISTRY

In good shape

J. Am. Chem. Soc. **127**, 17160–17161 (2005)

Widely used for knitting organic molecules together, the metathesis reaction netted its developers the 2005 Nobel Prize in Chemistry. But the carbon–carbon double bond in the reaction's product can sometimes end up in the wrong position.



M. WILDWATER ET AL. CELL

Robert Grubbs, who shared the prize, and his team at the California Institute of Technology in Pasadena find that adding traces of benzoquinone derivatives can almost entirely block the unwanted bond move.

They use the technique to make α -olefins. These have a double bond at the end of the molecule, where it is at risk of adopting the wrong position. The team is now investigating industrial applications of the mild, cheap benzoquinone additives.

DEVELOPMENT

Boy or girl?

Nature Neurosci. doi:10.1038/nn1624 (2006)

The advent of a mouse mutant lacking an oestrogen-binding protein has settled a long-standing debate over the role oestrogens play in the developing brains of the different sexes.

In the mouse fetus (pictured right) as in all mammals, the brain develops as male in the presence of oestrogens, which are synthesized within the brain from testosterone produced in fetal testes. But it was not known whether normal development of the female brain requires the absence of oestrogens.

Two opposing theories have been developed, centring on the role of the fetal blood protein, α -fetoprotein, which binds tightly to oestrogens. One holds that the protein blocks uptake of oestrogens, the other that it actively transports the hormone into the brain.

By analysing the brains and behaviour of mice lacking the α -fetoprotein gene, Julie Bakker from the University of Liège and her colleagues clearly show that prenatal oestrogens masculinize and defeminize the brain, and that α -fetoprotein protects the female brain from the effects of oestrogens.

SENSORS

As thick as blood

J. Am. Chem. Soc. doi:10.1021/ja056370a (2005)

Changes in the viscosity of body fluids can signify disease. But mechanical devices do not monitor such changes closely enough to provide a reliable diagnosis. That has led researchers from the University of Missouri–Columbia and the University of California, San Diego, to develop molecular viscosity sensors whose fluorescent emission

depends on how free the structure is to rotate.

Now work has built on this to overcome the problem of calibrating the sensor for emission intensity, which is influenced by sensor concentration and body-fluid properties. It presents a molecule that contains a built-in 'reference' emitter. Changes in viscosity alter the brightness of the sensor relative to the reference point, irrespective of other factors.

ASTRONOMY

Planetary ingredients

Astron. Astrophys. **445**, 633–645 (2005)

Many studies suggest that stars with planets are richer in heavy metals, such as iron, than their solitary neighbours. But it is unclear whether this is a cause or a consequence of their planetary systems.

To find out, Alexandra Ecuivillon of the Astrophysics Institute of the Canary Islands and her colleagues measured the oxygen content of 155 solar-type stars, 96 of which have planetary-mass companions. If stars with planets got their extra iron from planetary matter, their oxygen content should seem low by comparison. But the team found no clear difference in the ratio of oxygen to iron between the two groups of stars.

The finding adds to growing evidence that planetary systems are more likely to condense from iron-rich clouds.

GENOMICS

Variety is the dice of life

PLoS Genet. **1**, e78 (2005)

One of the things that helps make us all different is the way non-coding areas of our genome control the activity of our genes. Variations in these areas can influence our susceptibility to disease.

Seeking such differences, a group led by Panagiotis Deloukas and Emmanouil Dermitzakis at the Wellcome Trust Sanger Institute in Hinxton, UK, looked at bits of the genomes of 60 unrelated people. These genomes had already been studied as part of the International HapMap Project, which looks for differences in single genetic letters known as single nucleotide polymorphisms (SNPs).

The team found a surprisingly large amount of variation in gene activity between the people, and linked 40 examples to particular SNPs.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

JOURNAL CLUB

Seong-Seng Tan

Howard Florey Institute,
University of Melbourne,
Australia

A neuroscientist sets the brain's cells to music.

If the neurons in the brain were the world's great tenors, I'd argue that, like Luciano Pavarotti, projection neurons hog the limelight, whereas interneurons are the Plácido Domingos of the cortical stage.

Projection neurons are seen as more glamorous because of their greater number. They are also well connected inside and outside the brain, carrying messages through powerful axonal cords to distant parts of the body.

By contrast, interneurons have short, locally connecting axons whose main function seems to be suppressing excessive neuronal activity. This puts interneurons in their counterpart's shadow.

But many researchers think the synchronized humming of interneurons is essential to the richness of cortical processing. What makes their music underappreciated is the problems we have comprehending their repertoire.

Interneurons are hard to study, as they exist in a bewildering diversity. After accounting for different morphological, electrophysiological and chemical characteristics, there is a cast of thousands.

Recently, there has been some progress in understanding how different types of interneurons are made. Two groups have established correlations between the birthplace and birth date of a neuron, and its anatomy, physiology and molecular expression (S. J. Butt *et al.* *Neuron* **48**, 591–604; 2005 and Q. Xu *et al.* *J. Neurosci.* **24**, 2612–2622; 2004).

These results show that interneuron identity is specified early. It seems likely that interneurons have to read a libretto of transcription factors in their germinal zones before singing out their type. If we identify these factors, we can study the notes that define the interneurons' opera and maybe that will bring their music the recognition it deserves.

NEWS

Mashups mix data into global service

Will 2006 be the year of the mashup? Originally used to describe the mixing together of musical tracks, the term now refers to websites that weave data from different sources into a new service. They are becoming increasingly popular, especially for plotting data on maps, covering anything from cafés offering wireless Internet access to traffic conditions. And advocates say they could fundamentally change many areas of science — if researchers can be persuaded to share their data.

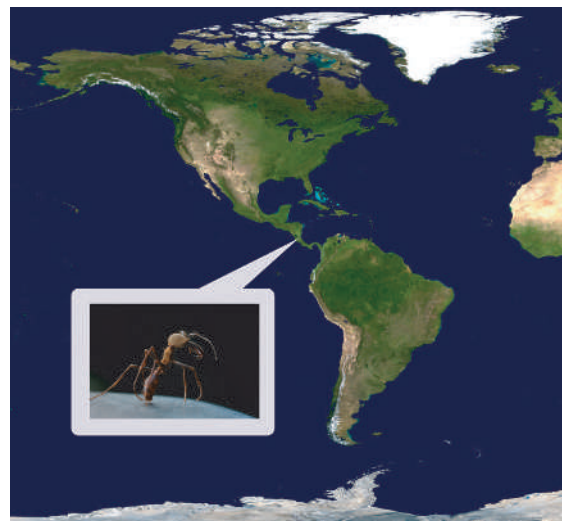
Some disciplines already have software that allows data from different sources to be combined seamlessly. For example, a bioinformatician can get a gene sequence from the GenBank database, its homologues using the BLAST alignment service, and the resulting protein structures from the Swiss-Model site in one step. And an astronomer can automatically collate all available data for an object, taken by different telescopes at various wavelengths, into one place, rather than having to check each source individually.

So far, only researchers with advanced

programming skills, working in fields organized enough to have data online and tagged appropriately, have been able to do this. But simpler computer languages and tools are helping.

Google's maps database, for example, allows users to integrate data into it using just ten lines of code (www.google.com/apis/maps). UniProt, the world's largest protein database, is developing its existing public interfaces to protein sequence data to encourage outside users to access and reuse its data.

The biodiversity community is one group working to develop such services. To demonstrate the principle, Roderic Page of the University of Glasgow, UK, built what he describes as a "toy" — a mashup called Ispecies.org (<http://darwin.zoology.gla.ac.uk/~rpage/ispecies>). If you type in a species name it builds a web page for it showing sequence data from GenBank, literature from Google Scholar and photos from a Yahoo image search. If you could pool data from every museum or lab in the world, "you could do amazing things", says Page.



Web crawling: ant researchers are bringing together information from a variety of sources.

Donat Agosti of the Natural History Museum in Bern, Switzerland, is working on this. He is one of the driving forces behind AntBase and AntWeb, which bring together data on some 12,000 ant species. He hopes that, as well as searching, people will reuse the data to create phylogenetic trees or models of geographic distribution.

This would provide the means for a real-time, worldwide collaboration of systematists, says Norman Johnson, an entomologist

M. DOHRN/K. TAYLOR/NATUREPI/NASA

Intelligent design verdict set to sway other cases

WASHINGTON DC

A high-profile trial centred on the teaching of evolution is over. High-school students in Dover, Pennsylvania, will not now hear an announcement promoting intelligent design — the idea that an intelligent creator shaped today's organisms — before taking lessons on evolution. On 20 December, federal judge John Jones struck down a local school-board decision in a scathing 139-page rebuke to the intelligent-design movement. But other challenges to evolution are simmering across the country — and the Dover decision could influence their outcome, some say.

Such fights usually originate at the state level — in the form of legislation or the setting of state-wide education standards — or at the school-district level, where local standards and curricula are generally set.

A recent study from the Washington-based Thomas B. Fordham Institute into science curriculum standards gave failing grades to 15 states (see map). Alabama students, for

instance, learn from biology textbooks adorned with a sticker describing evolution as "controversial". But in Ohio, which passed, some students are taught from a state-approved lesson plan called "critical analysis of evolution", in which they research and present pro- and anti-evolution viewpoints.

Robin Hovis, a member of the Ohio state board of education, says the Dover case may affect the future of the lesson plan. "It certainly gave those of us on the board who objected renewed hope," he adds. In Cobb County, Georgia, an appeals court is set to rule on a lower-court judgement deeming similar stickers unconstitutional.

And in Kansas, a school-board primary election next August could reshape the state's educational landscape. Board members who edited the education standards to include "scientific criticisms" of evolution face challenges by moderate Republicans who want such language weeded out (see *Nature* 438, 267; 2005).

In another twist, a group of Christian schools is suing the University of California for refusing to recognize certain high-school courses, including biology classes that use textbooks taking an anti-evolution view. The university has filed for dismissal, and expects to hear from the judge in a few months. "The Dover verdict says schools can't teach these non-scientific ideas as science, so that supports us," says Christopher Patti, a lawyer with the university.

Legislation promoting intelligent design or similar anti-evolution ideas was introduced in more than a dozen states in 2005. Most died a hasty death, according to Nick Matzke, spokesman for the National Center for Science Education, a California-based non-profit organization that fights for evolution education. He and others hope that the Dover decision will help quash the promotion of intelligent design, which they say is a legal strategy for introducing religion into the classroom. "Court decisions


**WHATEVER
HAPPENED TO...**

News@nature.com recalls
some top stories of 2005
to see how they turned out.
www.nature.com/news



at Ohio State University in Columbus. "It has the potential to fundamentally change and improve the way that basic systematic research is conducted."

A major limiting factor is the availability of data in formats that computers can manipulate. To develop AntWeb further, Agosti aims to convert 4,000 papers into machine-readable online descriptions. Another problem is the reluctance of many labs and agencies to share data.

But this is changing. A spokesman for the Global Health Atlas from the World Health Organization (WHO), for example, a huge infectious-disease database, says there are plans

to make access easier. The Global Biodiversity Information Facility (GBIF) has linked up more than 80 million records in nearly 600 databases in 31 countries. And last month saw the launch of the International Neuroinformatics Coordinating Facility.

But such initiatives are hampered by restrictive data-access agreements. The museums and labs that provide the GBIF with data, for example, often require outside researchers to sign online agreements to download individual data sets, making real-time computing of data from multiple sources almost impossible.

Nature has created its own mashup, which integrates data on avian-flu outbreaks from the WHO and the UN Food and Agriculture Organization into Google Earth (www.nature.com/nature/googleearth/avianflu1.kml). The result is a useful snapshot, but illustrates the problem. As the data are not in public databases that can be directly accessed by software, we had to request them from the relevant agencies, construct a database and compute them into Google Earth. If the data were available in a machine-readable format, the mashup could search the databases automatically and update the maps as outbreaks occur. Other researchers could also mix the data with their own data sets.

Page and Agosti hope that researchers will soon become more enthusiastic about sharing. "Once scientists see the value of freeing-up data, mashups will explode," says Page. ■

Declan Butler

Croatian scientists call for openness over funding

More than 250 scientists inside and outside Croatia have signed a petition calling for more transparency in the country's funding of science and technology.

In particular, the petition calls for an investigation into a technology-development grant issued two years ago to Dragan Primorac, now the country's science minister.

The scientists allege "irregularities" and possible conflicts of interest in the operation and funding of the €1-million (US\$1.2-million) project to establish a forensic and molecular-genetics laboratory — accusations Primorac denies. Primorac won the grant in December 2003, when he was director of a clinic at the Holy Spirit Hospital in Zagreb. He was appointed minister four days later.

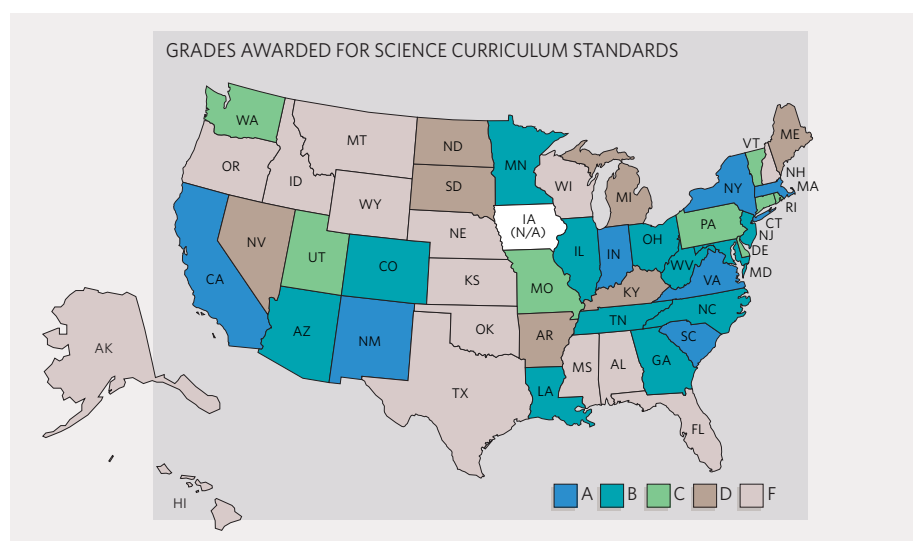
The petition was prompted by some scientists' unhappiness that, over a few months, the lab's grant money was transferred between several different institutes in Zagreb. They also want to know how Primorac resolved potential conflicts of interest relating to the grant — for example, he was a member of the technology council that evaluated and approved his own grant.

Vlatko Silobrcic, former director of the Institute of Immunology in Zagreb and a member of the Croatian Academy of Sciences, was one of the petition's 15 original signatories. He says that the episode is symptomatic of a general lack of openness in the way science money is allocated in the country. "The action was started to get answers to issues that we think are important for the science system in Croatia," he says. "It is legitimate to ask questions about possible conflicts of interest."

Krešimir Pavelić, director of molecular medicine at the Rudjer Bošković Institute, Croatia's largest research institute, is currently in sole charge of the project. An institute spokesman says that the moves were largely an attempt to find enough lab space.

Meanwhile, Primorac told *Nature* that he has always been open about the moves, and that concerns about conflicts of interest are unfounded. "I have nothing to hide," he says, adding that the criticisms are part of a "relentless campaign" against him. ■
Alison Abbott

THOMAS B. FORHAM INST.



never resolve social issues, and it won't here," says Matzke. "But it will give us a little breathing space. Intelligent design as a strategy is probably toast."

Naturally, proponents of the theory disagree. Casey Luskin, a lawyer at the

Discovery Institute, an intelligent-design think-tank in Seattle, Washington, says the Dover decision will have a "negligible effect". "You cannot change the facts of biology through a judicial ruling," he argues. ■
Emma Marris

Blow follows blow for stem-cell work

There is no evidence that cloning researcher Woo Suk Hwang created any lines of patient-specific stem cells. The verdict, reached by investigators at Seoul National University (SNU) after claims that Hwang faked his work, confirms stem-cell researchers' worst fears.

In a celebrated paper published last May, the South Korean researcher reported cloning cell lines from 11 different patients, which could in principle have been used for therapy (W. S. Hwang *et al. Science* **308**, 1777–1783; 2005). But after several allegations of fraud, a university committee asked three different laboratories to compare the DNA of Hwang's purportedly cloned cell lines with that of the patients who had provided the cells.

On 23 December, the committee reported that the data from the *Science* paper came from just two cell lines, not eleven as claimed. And on 29 December, it demolished hopes that those two lines would be patient-specific: they turned out to match stem-cell lines from embryos produced by *in vitro* fertilization at MizMedi Hospital in Seoul. MizMedi provided egg cells for Hwang's research.

"The committee concludes that there are neither patient-specific stem-cell lines in Hwang's laboratory nor any scientific evidences to support the claim that such cell lines ever existed," said Jung Hye Roe, director of research at the SNU.

According to the Korean press, Hwang has accused a co-author at MizMedi of switching the cell lines. But he submitted his resignation

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

In deep: a university panel found no evidence that Woo Suk Hwang created patient-specific lines.

to the SNU on 23 December, and he and several co-authors have asked for the *Science* paper to be retracted.

On 10 January, the SNU plans to report on two other papers by Hwang's team — a *Science* paper in which the team claimed to be the first to produce stem cells from a cloned human embryo (W. S. Hwang *et al. Science* **303**, 1669–1674; 2004), and a *Nature* paper that reported the first cloned dog (B. C. Lee *et al. Nature* **436**, 641; 2005). Both journals are running investigations.

And a South Korean bioethics committee will meet on 15 January to look into the charges and other issues concerning egg donation in Hwang's lab. For example, Hwang claimed that he used only 242 eggs for the 2004 paper and 185 for the paper on patient-specific cells. But reports have emerged that say Hwang used more than 1,600 eggs in the experiments. Sung Il Roh, a MizMedi fertility expert and second author on the 2005 *Science* paper, told *Nature* that he alone provided 313 mature eggs, from 21 donors, for the 2004 paper and 900 eggs, from 62 donors, for the 2005 paper.

As *Nature* went to press, *PD Notebook*, a Korean investigative news programme, was about to air a piece that levels further charges against Hwang. According to one of its producers, the programme will claim that Hwang coerced a junior team member into donating eggs for research. *Nature* previously reported that this researcher felt obliged to donate eggs after accidentally destroying some early in the experiment. Hwang has not so far responded to *Nature's* requests for an interview.

Hwang might even find himself facing legal action. According to *The Korea Times*, government prosecutors could decide to pursue fraud charges after the SNU investigation is complete. Hwang's group received more than 3 billion won (US\$3 million) for research from the ministry of science and technology last year, and 24 billion won for new facilities. ■

David Cyranoski

LEE JIN-MAN/AP

Los Alamos bosses ditched after decade of scandals

WASHINGTON DC

Los Alamos National Laboratory will be under fresh management from June. But staff wonder whether their new bosses will help the lab move on from the scandals that have dogged it in the past decade.

On 21 December, the US Department of Energy announced that the University of California would continue to run the New Mexico institution as part of a consortium that also includes three private companies. In a competition for the lab, it beat a partnership between the University of Texas and defence contractor Lockheed

Martin. Los Alamos has been run by the University of California since it was set up to develop the atomic bomb during the Second World War.

The competition was the first in the lab's history, and was triggered by a series of security scandals, including charges of espionage, an industrial accident and reportedly missing computer disks that caused a three-month shutdown in 2004 (see *Nature* **433**, 437; 2005).

In announcing the management change, energy department officials were quick to reassure employees who are tired of the recurrent problems. "I cannot stress enough

that this is a new contract, with a new team, marking a new approach at Los Alamos," said Samuel Bodman, the US energy secretary.

Along with the University of California, the consortium — called Los Alamos National Security LLC — consists of Bechtel, BWX Technologies and Washington Group International, which are all nuclear-facilities contractors. Some employees think that it will therefore concentrate on nuclear weapons at the expense of basic science. "There is a lot of concern that Los Alamos would become nothing more than a plutonium

factory," says Doug Roberts, who recently retired from Los Alamos and runs a widely read blog about the lab.

Michael Anastasio, the lab's new director, has tried to allay such concerns. "This is not a de-emphasis on science," says Anastasio, currently director of the Lawrence Livermore National Laboratory in California.

Under the new set-up, the university is expected to run science operations, while the other partners take care of administrative matters, including security. ■ Emma Marris

Science clings on as Japan slashes spending

Despite an overall cut of 3% to Japan's national budget for 2006, basic research managed to secure a slight increase in funds.

Approved by the cabinet on 24 December, the budget includes ¥1.3 trillion (US\$11.4 billion) for the country's main research activities, a rise of 1.1% on last year. This was welcome news for scientists, who had been steeling themselves for cutbacks (see *Nature* 437, 181; 2005).

Among the winners are the Japan Aerospace Exploration Agency, which received a 2% much-needed boost to ¥180 billion, and a bid to build a fast-breeder nuclear reactor, which scored a 13% increase to ¥34.7 billion. New projects aimed at developing the next generation of supercomputers and a powerful X-ray free-electron laser for atomic-scale imaging will get ¥3.5 billion and ¥2.3 billion, respectively.

But not all sectors of science are so fortunate. The life-sciences budget will fall by 11% to ¥74.4 billion, largely because a project for the International Space Station was cancelled. The government is also cutting back subsidies for public universities and research institutes, while encouraging them to bid for competitive grants and to seek collaborations with private companies.

Death sentence revoked for HIV workers in Libya

Six health workers sentenced to death in Libya were granted a reprieve on 25 December. The country's supreme court quashed the sentence and ordered a retrial.

In 1999, five Bulgarian nurses and a Palestinian doctor were accused of deliberately infecting 426 Libyan children with HIV. They insist they are innocent and that their confessions were extracted under torture. Investigations by independent experts have concluded that the children were probably infected through unsanitary hospital conditions with the outbreak beginning before the health workers arrived in the country (see *Nature* 430, 277; 2004).

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Libya's supreme court has revoked the death sentence passed on six foreign health workers.

Upgrade aims to put drill ship at cutting edge

After 20 years as the flagship of scientific ocean drilling, the *JOIDES Resolution* (right) is taking a holiday. The ship is heading to dry dock for what its operators describe as an "extreme makeover".

In a US\$115-million facelift, the ship will get a new name, at least 50% more lab space and revamped drilling systems to enhance its sample retrieval and analysis capabilities.

Creature comforts will also be improved with the addition of a sauna. And accommodation will see just two people sharing a berth rather than the current four.

Once operational, in summer 2007, the revamped ship is expected to join Japanese-built drillship *Chikyu* off the coast of Japan.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Together, as part of the Integrated Ocean Drilling Program, they will try to drill into an active fault in the crust, some 5.5 kilometres beneath the ocean floor.

Bulgaria, Libya, Britain, the European Commission and the United States recently agreed a plan to provide money for the infected children's medical care — a move that some think led to the overturning of the death sentences.

Europe launches first part of navigation network

The first test satellite in the Galileo network, the European equivalent of the US Global Positioning System (GPS), launched on 28 December.

Costing €3.8 billion (US\$4.5 billion), Galileo's full constellation of 30 satellites is expected to be operational in 2008, joining the 29 satellites in the GPS network. Researchers plan to use both Galileo and GPS for a range of projects — from measuring the creeping movement of Earth's tectonic plates to the water content of the lower atmosphere.

"GPS is still primarily a military tool," says Terry Moore, a satellite-navigation expert at the University of Nottingham, UK; its signals could in theory be switched off at any time. Galileo, on the other hand, is primarily for civilian use and should stay on in all but the most extreme circumstances.

Terrorist attack in India raises alarm for scientists

Scientists found themselves the target of a terrorist attack in India late last year. On 28 December, gunmen shot and killed retired mathematics professor Munish Chandra Puri at the Indian Institute of Science (IISc) in Bangalore. As a result, the Indian Science Congress, the country's largest scientific event, opened on 3 January in Hyderabad amid unprecedented security.

Puri, who was based at the Indian Institute of Technology in Delhi, was killed in what officials said was a senseless shooting designed to create panic across the scientific community. Four others were injured, including the IISc's Vijay Chandru, developer of a cheap hand-held computer.

The attackers picked scientists as soft targets, says science secretary Valangiman Ramamurthy, and they chose Bangalore — home to about 1,500 software companies — to attract global attention. A man belonging to the Pakistan-based militant group Lashkar-e-Taiba, which wants to see a separate Kashmir, has been arrested in connection with the shooting. Several others have been detained.

Freed archaeologist considers return to Iraq

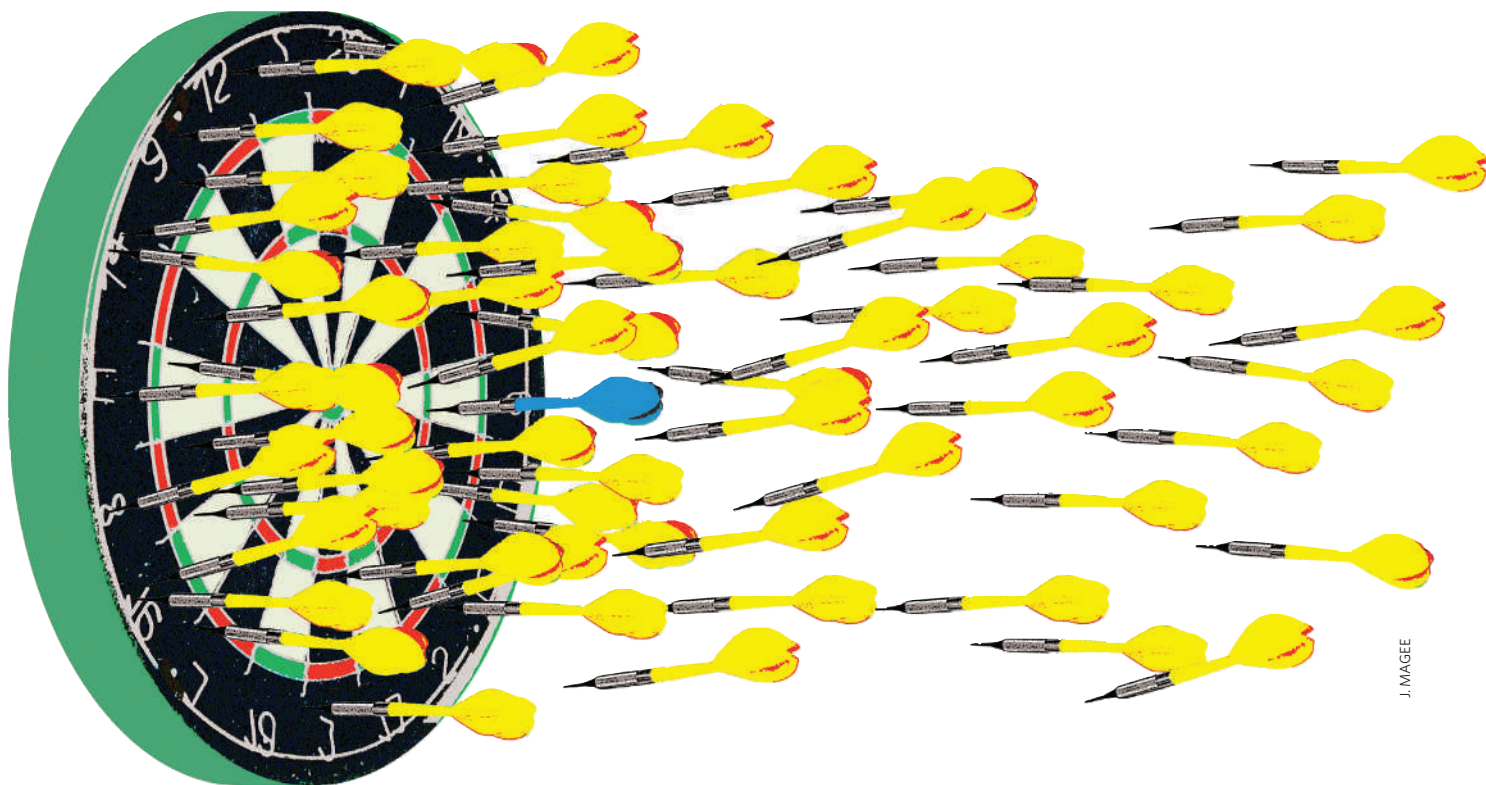
German officials are urging archaeologist and aid worker Susanne Osthoff not to return to Iraq, where she was kidnapped and held for three weeks before being released on 18 December.

Osthoff has been trying for years to draw attention to the looting of archaeological sites (see *Nature* 438, 722; 2005). She has also organized the transport of medicine and other goods to Iraqi citizens.

In television interviews after her release, Osthoff left it open as to whether she will return to Iraq. Her comments have angered some German politicians, who say that the country should not fund future rescue efforts if Osthoff is kidnapped again.

The German government has also refused to continue funding her projects in Iraq, including the renovation of a historic building in Mosul on which it has already spent €40,000 (US\$47,500).

It is still unclear who kidnapped Osthoff and how her release was secured.



J. MAGEE

OUTRAGEOUS FORTUNE

A growing number of cosmologists and string theorists suspect the form of our Universe is little more than a coincidence. Are these harmless thought experiments, or a challenge to science itself? **Geoff Brumfiel** investigates.

Why are we here? It's a question that has troubled philosophers, theologians and those who've had one drink too many. But theoretical physicists have a more essentialist way of asking the question: why is there anything here at all?

For two decades now, theorists in the think-big field of cosmology have been stymied by a mathematical quirk in their equations. If the number controlling the growth of the Universe since the Big Bang is just slightly too high, the Universe expands so rapidly that protons and neutrons never come close enough to bond into atoms. If it is just ever-so-slightly too small, it never expands enough, and everything remains too hot for even a single nucleus to form. Similar problems afflict the observed masses of elementary particles and the strengths of fundamental forces.

In other words, if you believe the equations of the world's leading cosmologists, the probability that the Universe would turn out this

way by chance are infinitesimal — one in a very large number. "It's like you're throwing darts, and the bullseye is just one part in 10^{120} of the dart board," says Leonard Susskind, a string theorist based at Stanford University in California. "It's just stupid."

One in a zillion

Physicists have historically approached this predicament with the attitude that it's not just dumb luck. In their view, there must be something underlying and yet-to-be-discovered setting the value of these variables. "The idea is that we have got to work harder because some principle is missing," says David Gross, a Nobel-prizewinning theorist and director of the Kavli Institute for Theoretical Physics in Santa Barbara, California.

But things have changed in the past few years, says astronomer Bernard Carr of Queen Mary, University of London, UK. String theorists and cosmologists are increasingly turning to dumb luck as an explanation. If their ideas

stand up, it would mean the constants of nature are meaningless. "In the past, many people were almost violently opposed to that idea because it wasn't seen as proper science," Carr says. "But there's been a change of attitude."

Much of that change stems from work showing that our Universe may not be unique. Since the early 1980s, some cosmologists have argued that multiple universes could have formed during a period of cosmic inflation that preceded the Big Bang. More recently, string theorists have calculated that there could be 10^{500} universes, which is more than the number of atoms in our observable Universe. Under these circumstances, it becomes more reasonable to assume that several would turn out like ours. It's like getting zillions and zillions of darts to throw at the dart board, Susskind says. "Surely, a large number of them are going to wind up in the target zone." And of course, we exist in our particular Universe because we couldn't exist anywhere else.

It's an intriguing idea with just one problem,

says Gross: "It's impossible to disprove." Because our Universe is, almost by definition, everything we can observe, there are no apparent measurements that would confirm whether we exist within a cosmic landscape of multiple universes, or if ours is the only one. And because we can't falsify the idea, Gross says, it isn't science. Or at least, it isn't science in any conventional sense of the word. "I think Gross sees this as science taking on some of the traits of religion," says Carr. "In a sense he's correct, because things like faith and beauty are becoming a component of the discussion."

And yet in the overlapping circles of cosmology and string theory, the concept of a landscape of universes is becoming the dominant view. "I really hope we have a better idea in the future," says Juan Maldacena, a string theorist at the Institute for Advanced Study in Princeton, New Jersey, summing up the views of many in the field. "But this idea of a landscape is the best we have today." The stakes are high: string theorists know that pursuing an unverifiable theory could look like desperation, but they fear that looking for meaning in a meaningless set of numbers may be equally fruitless.

Kepler's error

At the core of this dilemma is a concept known as the anthropic principle: the idea that things appear the way they do because we live at a certain spot in the Universe. It's not a new concept, and has previously been regarded more as philosophy than science.

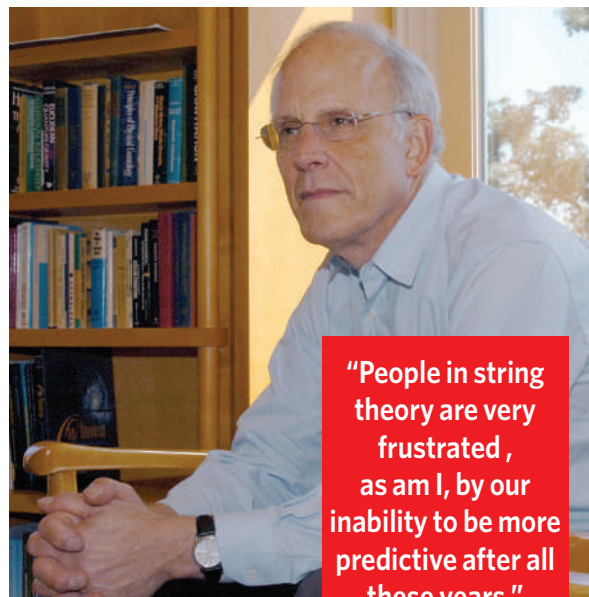
But some scientists say that it offers a useful change of perspective. "It's very important to take into account stuff like this, or you can

come to completely incorrect conclusions about the Universe," argues Max Tegmark, a cosmologist at the Massachusetts Institute of Technology, Cambridge. "For example, you might assume our Solar System is typical, but a typical point in space is some intergalactic void where you can't see a single star."

Failing to consider our observational location has burned scientists in the past. The sixteenth-century German astronomer Johannes Kepler spent years trying to understand what seemed to be the even, geometrical spacing of our planets from the Sun. Kepler searched for meaning in the planets because he thought our Solar System was unique; today's scientists understand that our Solar System is but one of probably billions in the Galaxy. Under such circumstances it seems reasonable to assume the planets are spaced according to little more than random chance.

In much the same way as Kepler worried about planetary orbits, cosmologists now puzzle over numbers such as the cosmological constant, which describes how quickly the Universe expands. The observed value is so much smaller than existing theories suggest, and yet so precisely constrained by observations, that theorists are left trying to figure out a deeper meaning for why the cosmological constant has the value it does.

Many are still searching for some great unifying theory that would explain these variables. But others have started to believe that, like Kepler, today's physicists are looking for meaning where there is none. "In recent years, it was looking more and more to me like the laws of nature were environmental," says Susskind, who has just written a book making this argument (*L. Susskind The Cosmic Landscape: String Theory and the Illusion of Intelligent Design*. Little Brown, 2005). He suspects that there are many universes, all with different values for these variables. Just as human



"People in string theory are very frustrated, as am I, by our inability to be more predictive after all these years."

DAVID GROSS

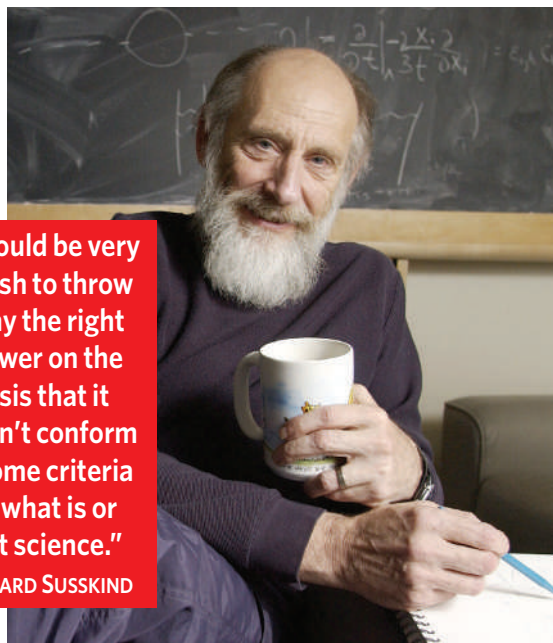
life had to evolve on a planet with water, he says, perhaps we also had to evolve in a Universe where atoms could form.

Until recently, Susskind was in the minority. Hints of multiple universes, however, were given by a cosmological theory known as inflation. Inflation is the leading theory of the early Universe; it postulates that a period of rapid early expansion created the flat and uniform cosmos we see today. One version of inflation theory, devised in the early 1980s, suggests that inflation occurred even before the Big Bang. In this version, the expanding cosmos was foamy and energetic, says Steven Weinberg, a researcher at the University of Texas, Austin. "Every once in a while, one part of the Universe would expand and become a Big Bang," he says. "And these Big Bangs would all have different values for their fundamental constants."

Strings attached

In 1987, Weinberg made a prediction that turned out to support the idea of an anthropic Universe. Preliminary observations indicated that the cosmological constant was zero, but Weinberg reasoned that if the constant was constrained by our anthropic perspective then it would be small, so as not to interfere with the formation of galaxies, stars and planets, but non-zero, because it would be essentially random. "That prediction has since been confirmed by observations of supernovae and the microwave background," says Weinberg, who admits he was a reluctant convert to the idea.

The latest circumstantial arguments for multiple universes come from string theory.

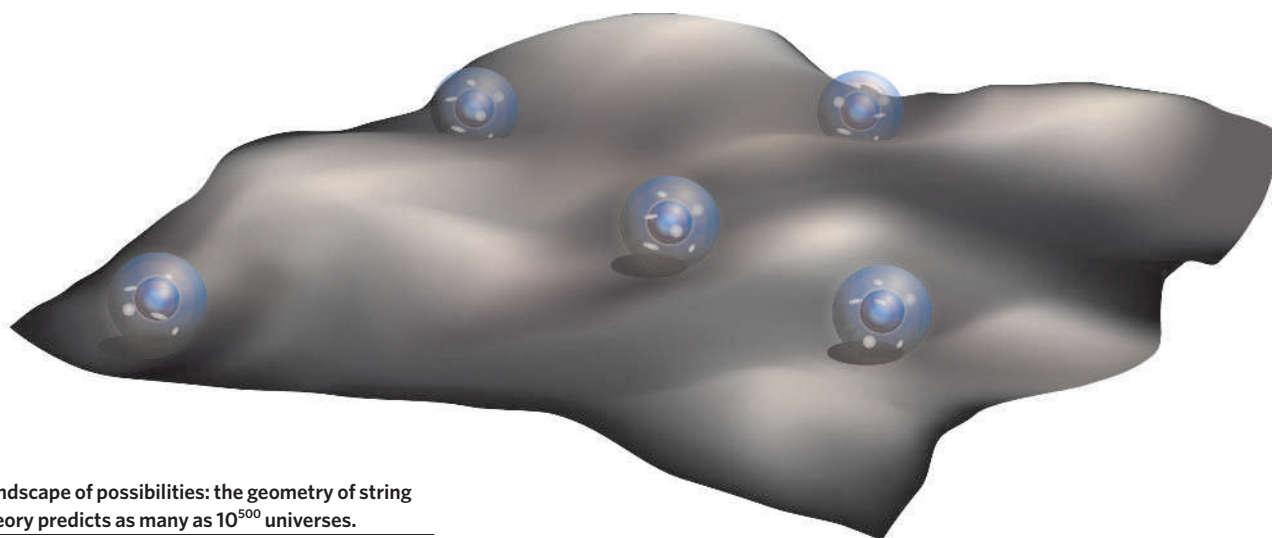


"It would be very foolish to throw away the right answer on the basis that it doesn't conform to some criteria for what is or isn't science."

LEONARD SUSSKIND

P. KLEIN/REUTERS

L. CICERO/STANFORD NEWS SERVICE



Landscape of possibilities: the geometry of string theory predicts as many as 10^{500} universes.

String theory posits that tiny strings vibrating in the fabric of space-time give rise to the multitude of particles and forces in the macroscopic Universe. Although string theory lacks experimental support, it attracts broad interest because it seems to offer a route to a grand theory of everything — a way to unify relativity with quantum mechanics.

But as theorists developed string theory, they discovered that the equations gave rise to multiple solutions, each of which represented a universe with different physical properties. “The hope always was that we would understand why one solution was picked out,” says Joe Polchinski, a string theorist at the Kavli Institute. But despite their best efforts, after two decades theorists are still stuck with a million different solutions for the equations, and therefore a million potential universes.

This landscape of solutions, as it became known in the community, was both troubling and intriguing. On the one hand, the theory stubbornly refused to yield a single solution resembling our own cosmos, but then, some argued, that might also explain the cosmological constant’s apparent randomness. If these many solutions actually represent millions of universes, then the idea that one had worked out just right for us wasn’t so far-fetched.

Ignorance is bliss

The snag was that one million universes wasn’t enough. To explain the perfectly adjusted cosmological constant one would need at least 10^{60} universes, says Polchinski. Then, in 2000, he and Raphael Bousso at the Lawrence Berkeley National Laboratory in Berkeley, California, calculated that there could be a lot more than a million solutions. “The calculation had such topological complexity that you could potentially get 10^{500} universes,” Polchinski says. With so many solutions, says Weinberg, it becomes easier to imagine that we happen to live in a Universe that seems tailored for our existence.

Easy to imagine, hard to prove. Because other universes would be causally separated

from our own, it seems impossible to tell whether our cosmos is the only one, or one of many. Most scientists find this disturbing. Talk of a Universe fine-tuned for life has already attracted supporters of intelligent design, who claim that an intelligent force shaped evolution. If there’s no way to tell whether the values of scientific constants are a coincidence, the movement’s followers argue, then why not also consider them evidence of God’s handiwork?

The anthropic reasoning behind the landscape of universes is disturbing on another level, says Gross. Most theories grow stronger with each observation that matches their predictions. However, for the anthropic principle, random chance is the main factor. Patterns and correlations, the stones from which scientific theories are built, weaken it. In other words, he says: “The power of the principle is strongest where you have ignorance.”

That may be, but measurements that could support anthropic reasoning are in the works. In 2007, researchers at Europe’s CERN particle physics centre in Geneva, Switzerland, will turn on the Large Hadron Collider, a massive accelerator that will probe particle energies never before seen by researchers. The accelerator might detect so-called supersymmetric particles, predicted by some as a way of unifying the strong and weak nuclear forces with the electromagnetic force, an important step in uniting all the forces of physics within a single theory.

These particles could also hint at whether we live in one of many universes, says Nima Arkani-Hamed, a string theorist at Harvard University in Cambridge, Massachusetts. If the collider detects certain types of supersymmetric particles, he says, it will indicate another fine-tuning in the cosmos — the ratio of the weak nuclear force to the strength of gravity. The anthropic argument is the same: if the number was off by as little as one part in 10^{30} , then we would not be here to discuss it.

It might seem that the detection of a second,

perfectly tuned number would only exacerbate the debate, but Arkani-Hamed argues that it will have the opposite effect. Unlike the cosmological constant, which has had a controversial history even in cosmology, this fine-tuning would appear in the standard model, which most physicists consider to be the most complete physical theory ever developed and tested. It would strengthen the case for the arbitrary nature of certain fundamental constants, Arkani-Hamed contends: “These measurements wouldn’t directly prove or disprove the landscape, but they would be a very big push in that direction.”

Leap of faith

Still, many scientists distrust the concept and continue to seek alternative explanations. Among them is Lisa Randall, also at Harvard. Randall suspects that multiple universes are a mirage resulting from the unrefined equations of string theory. “You really need to explore alternatives before taking such radical leaps of faith,” she contends. And with no foreseeable way to detect other universes, Gross feels that such leaps of faith should not be taken. “I feel that it’s a rather extreme conclusion to reach at this point,” he says.

Susskind, too, finds it “deeply, deeply troubling” that there’s no way to test the principle. But he is not yet ready to rule it out completely. “It would be very foolish to throw away the right answer on the basis that it doesn’t conform to some criteria for what is or isn’t science,” he says.

Gross believes that the emergence of multiple universes in science has its origins in theorists’ 20-year struggle to explain the finely tuned numbers of the cosmos. “People in string theory are very frustrated, as am I, by our inability to be more predictive after all these years,” he says. But that’s no excuse for using such “bizarre science”, he warns. “It is a dangerous business.”

Geoff Brumfiel is Nature’s physical sciences Washington correspondent.

AWASH WITH FOSSILS

The Afar region of Ethiopia is littered with traces of the earliest humans. **Rex Dalton** gets on the trail with a team of devoted experts who just live for the next find.

R. DALTON



Into Africa: Berhane Asfaw has devoted two decades to unearthing Ethiopia's treasures.

Ethiopians have a peculiarly fitting saying about the remote province of Afar — it is 'where it all began'. Isolated in the harsh region is a low ridge that has brought to light humanity's beginnings, by providing a wealth of fossils.

In early December, under the glare of a near-equatorial sun, an international team of palaeoanthropologists scours sediments on the ridge for fossil bones, teeth or artefacts; the researchers have worked here for more than 20 years. Five years ago they identified a site called Asa Issie, or 'red hill', as promising ground after unseasonable torrents trapped them in the area. This year they have returned to see what has surfaced after summer rains swelled the nearby Awash river.

At times stooped, sometimes crawling, but always focusing on the sea of stones that cover the ground, the crew spreads slowly through ravines. Alongside the scientists are locals, whose sharp eyes and self-taught knowledge make them integral to success.

First come fossils of rhinoceros (*Ceratotherium*) and kudu (*Tragelaphini*), bones that provide insight into the region's former environment. Then a shout pierces the afternoon wind: "Canine!"

A cream-coloured hominid tooth, about 4 million years old, has been spotted among the white pebbles. It is only the third day of a month-long field season, and this discovery marks the 13th consecutive year that the team has found hominid specimens. Later, as the sun sets over the mountains to the west, team co-leader Tim White says to me: "There is no other spot on the planet like this. It is a special place."

There is also no other team like White's. Year after year, the Middle Awash project has identified some of the most important hominid specimens, including one of the oldest *Homo sapiens* and the 5.8-million-year-old *Ardipithecus kadabba*¹⁻³. As other palaeoanthropologists have sought older and older hominids — and the fame that comes with such discoveries — the Middle Awash team has shone by identifying new hominid characteristics that flesh out the evolutionary track from ape to man.

Much of this success can be traced to the project's multinational roots. It represents the



best of scientific capacity-building: African scientists receive doctorates at top universities overseas, and then return to work and nurture projects at home. Scientists from abroad, such as White — at the University of California, Berkeley — are in the minority. The team's other leaders are Ethiopians: there is Giday WoldeGabriel, a geologist at the Los Alamos National Laboratory in New Mexico; the palaeoanthropologist Berhane Asfaw, director of Ethiopia's National Museum in Addis Ababa; and Yonas Beyene, a government archaeologist. Yohannes Haile-Selassie, a key team member who received his doctorate at Berkeley, like Asfaw, is curator of physical anthropology at the Cleveland Museum of Natural History in Ohio, a bastion of hominid research.

Postdocs and students come from France, Lebanon, Turkey and the United States. Everyone eats at the same table; team leaders shovel dirt just like the others. White drives the team hard, but junior scientists thirst for his feedback. One night, postdoc Michael Black — a specialist in hominid biomechanics who doubles as the camp solar electrician — waits so long for White's last instructions that he falls asleep at the dinner table.

The hominid tooth was found by Ferhat Kaya, a graduate of Turkey's Ankara University who studies minuscule mammals. Desperately working on his English skills to get on a US doctoral programme, Kaya beams when White gives him a complimentary high-five and tells him: "Good job, Ferhat." And everyone is pleased that the trip's first hominid discovery is made by the only trained scientist in the crew who is Muslim. "Fantastic," says Asfaw. "It shows the international aspect of the team."

Every year, the Middle Awash team ventures into the field in November or December, when the climate is most suitable for exploration. But rains still come. Once, a French researcher desperate to leave escaped by hiring Afar tribesmen to pull him on a raft to the asphalt highway.

Political upheavals have also threatened the project. Fieldwork stopped during most of the 1980s because of domestic turmoil and the writing of new antiquities laws. This year, David Perlman of the *San Francisco Chronicle* and I went to join the team, the first time journalists have been allowed to come along since its explorations began in 1981.

The Afar fossil ground is only about 250 kilometres northeast of Addis Ababa, which is located in the highlands near the centre of the country (see map). But the journey down to the Middle Awash site is a gruelling and dangerous three-day drive, and the crew has to make its own way along the final stretch on to a ridge known as Bouri.

Scene from Afar

The arid land is dotted by small villages of crude huts, often abandoned when the herders seek fresh grass. To me, the hamlets seem little changed from earlier eras, but today's Afar lifestyle might be better described as 'Neolithic, with Kalashnikovs'. Virtually every herder carries one of the Russian rifles, and some are equipped with rocket-propelled grenades.

The Afar are fighting another tribe, the Issa, which pushed north from Somalia seeking grazing lands. Daily, the palaeoanthropologists seek intelligence to avoid potential fire-fights. Armed policemen accompany the team at all times, as do Afar sheikhs and leaders, such as tribal chief Hamed Elema — who has himself become a skilled fossil finder. A year ago, the team had to abort a trip to date new hominid specimens because a gun battle threatened.

Last month, nine heavily loaded, all-wheel-drive trucks moved northeast out of Addis Ababa, through the Awash river valley and to the Bouri 'peninsula' — so named for its shape on the satellite images used to find sites. The drive features spectacular views of the mountains created by tectonic action along the African rift zone. The rift runs along the east of Africa; in the Afar region, the Arabian and Somalian plates pull east and away from the Nubian plate, causing giant blocks of crust to shift.

Standing on the Bouri ridge, looking west

toward the rift margin, the mountains resemble giant stairs: whole sections have dropped down after the tectonic expansion. To the north, the blocks have shifted and twisted, creating a complex geological terrain of ridges. To the south is Yardi Lake, a water-filled depression full of crocodiles and hippos.

Rain washes the edge of the mountainous blocks, exposing fossils of various ages. But dating the specimens is an onerous chore. It can take weeks of arduous hiking to collect samples for geochemical testing. The easiest and best way to find a fossil's age is to date a volcanic tuff layer above or below it. However, when the tuffs have been altered by weather and time, they produce no datable signature. Then geologists must correlate fossil-bearing layers with distant sediments whose ages are already known.

It took several years, for instance, for WoldeGabriel and Haile-Selassie to determine the age of the *A. kadabba* specimen⁴. "The geology is still ongoing," says WoldeGabriel. "I need to go and check some of the more complex faulting."

The remote treks put even Ethiopians in personal peril. In the late 1990s, when Haile-Selassie and WoldeGabriel were exploring the region near a sacred village, an offended Afar tribesman chased them away. "He kept his Kalashnikov on us," recalls WoldeGabriel. "It didn't make any difference that we were Ethiopians."

The work was worth it, though. Their 2001 article showed that the hominids — then the earliest known — lived in a wooded environment, not a savannah as previously thought.

On the Bouri peninsula, just a short drive is needed to reach interesting sediments. The early *H. sapiens* was found on the ridge, and a 1-million-year-old *H. erectus* was uncovered near the Bouri hamlet⁵, also called the 'hyena condominiums' because the beasts occupy the caves, bursting out when the team starts exploring. A 2.5-million-year-old *Australopithecus garhi* was discovered⁶ just where the ridge descends to the Awash river — a site that is known for extremely early evidence of stone-tool cuts on antelope bones. And just across the river is the Maka site, where the team discovered the 3.4-million-year-old *A. afarensis* that sparked one of its first major publications⁷.

Bouri is starkly different from the famous Hadar location, home to the 1974 discovery of the *A. afarensis* skeleton known as Lucy⁸. Just 70 kilometres to the north, Hadar is rich in hominid fossils — but they all come from the

"Back at the trucks, a girl waits with an elephant tooth. White jokingly suggests our police guards arrest her, and she is asked to return the fossil to its location."

R. DALTON



R. DALTON



Collector's lot: Tim White (right) has found many of the fossils that will grace the new palaeoanthropology wing of Ethiopia's National Museum.

same time period, 3 million to 3.5 million years ago. Bouri, in contrast, contains varied ages and species, which help palaeoanthropologists to understand the evolving features of man.

To sort through all this material, the Middle Awash team has devised sophisticated collection methods⁹. In one, 'the crawl', the researchers mark a promising area with surveyors' lines, then crawl across it shoulder to shoulder, removing every fossil found. Later comes 'maintenance', in which they return at various intervals to see what else has eroded from the plot — if fossils aren't found shortly after being exposed, they disintegrate.

With field time valuable, White is trying an experiment to determine the most efficient way to monitor sites. He planted 200 casts of fossils at an Awash site. The crew will return periodically for maintenance; by seeing how many casts are found, White hopes to determine how often return visits are needed if nothing is to be missed.

Such care underscores the long-term views of the team. White dislikes what he calls "hominid treasure hunts", where researchers move in for short field visits to grab hominids and then headlines.

For the team, the cataloguing of animal fossils deserves the same care as preserving hominid fossils. During last year's field season, the team collected about 1,400 vertebrate fossils — from elephant bones to the teeth of tiny mammals, which they preserve on waxed pinheads. All are cleaned and stored at the National Museum in Addis Ababa, which holds some 15,600 vertebrate fossils. The Ethiopian government is building a major new \$3.5-million research facility for the museum, and a whole wing will be set aside for the palaeoanthropology collection.

Such repositories help the study of hominids,

whichever team discovers them. For instance, French palaeoanthropologist Michel Brunet helped date the oldest hominid, *Sahelanthropus tchadensis*¹⁰, in Chad using a coexistent pig (*Nyanzachoerus syrticus*) from Ethiopia's collection. The pig species is known to have disappeared about 5.7 million years ago, so Brunet knew his specimen, called Toumaï, had to be at least that old.

Now, a former member of Brunet's team, Jean-Renaud Boissérie of Berkeley, is in the Middle Awash seeking to augment the animal fossil record — particularly that of hippopotamuses. These creatures can provide exquisite detail on dates and palaeobiogeography. They live in water, but species differ among river basins. Boissérie seeks hippo skulls and teeth, from which he can extract carbon-isotope samples and learn about the animals' diet and environment.

Tuff choices

"So what period do you want to examine?" White asks Boissérie on the drive to Bouri, offering him several options. Boissérie selects 2.5 million years. That means the team will head first to the Lubaka site near the Awash river, keeping eyes open for roaming lions.

Largely barren of vegetation, the hillocks at Lubaka reflect the volcanic tuffs used to date specimens. There is also another target today: stone tools. Although the *A. garhi* specimen was found with antelope bones that had cut marks, there were no artefacts. White and Asfaw want to link tools to the hominids.

After an hour and a half of hot walking, Boissérie gets his wish. Team member Kampire Krantu finds a good hippo jaw, with teeth, that is just surfacing in a ravine. Krantu is one of the team's best searchers. He has no scientific training, and no one on the team can

speak his Konso language. But his talent for spotting fossils is legendary.

The team leaders are not so lucky. White spots a black basalt chopper in a hillside. But he rejects it when he notices evidence in the sediments encasing it that suggest it has been washed there from its original location. Finding a tool that is truly *in situ* must wait for another day.

After a week in the field, the team cuts a new road out of the bush, making it easier for supply trucks to come and go. They begin to explore intriguing foothills south toward the rift margin. Hominid fever, however, dwindles when the only find is a giant tortoise fossil.

Back at the trucks, an Afar girl waits for us with a brick-shaped elephant tooth. White jokingly suggests our police guards arrest her, and she is asked to return the fossil to its location. He worries that keeping such items encourages locals to remove fossils from their surroundings, destroying vital geological information.

A week later, the team returns to ensure the fossil was replaced. And there, near the elephant tooth site, they find fresh fuel for their fever — a hominid tooth shard.

Bushels of earth around it will be sieved for any remaining pieces. Tuff dating will follow, then site maintenance. A published article may be years off, but once again Afar shows where it all began. ■

Rex Dalton is Nature's West Coast correspondent.

1. White, T. D. *et al. Nature* **423**, 742–747 (2003).
2. Haile-Selassie, Y. *Nature* **412**, 178–181 (2001).
3. Haile-Selassie, Y. *et al. Science* **303**, 1503–1505 (2004).
4. WoldeGabriel, G. *et al. Nature* **412**, 175–178 (2001).
5. Asfaw, B. *et al. Nature* **416**, 317–320 (2002).
6. Asfaw, B. *et al. Science* **284**, 629–635 (1999).
7. White, T. D. *et al. Nature* **366**, 261–265 (1993).
8. Johanson, D. C. & Taieb, M. *Nature* **260**, 293–297 (1976).
9. White, T. D. *C. R. Palevol* **3**, 341–351 (2004).
10. Brunet, M. *et al. Nature* **418**, 145–151 (2002).

BUSINESS

REX FEATURES

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Paint it green: John Browne has gone to great lengths to establish BP's environmental credentials.

More than just hot air?

With the launch of an alternative-energy division, BP is taking steps to show that it is serious about 'clean' technology. Emma Marris reports.

Back in 2001, BP shrugged off its full name of British Petroleum and began suggesting in its adverts that, if anything, the two letters stood for 'beyond petroleum'. Since then, the world's second-largest oil company has been working hard to promote itself as greener than its competitors.

Now it has channelled its efforts into a newly minted division: BP Alternative Energy. With 2,500 employees, this unit faces the task of rapidly expanding BP's clean-energy business. It will develop and sell technology that makes more efficient use of fossil fuels, as well as equipment for producing energy from renewable sources. In addition, the unit will use its technologies to run its own power plants.

BP says that in the run up to 2008 it will invest US\$600 million a year in the new unit, with more to come thereafter. It expects the division to generate revenues of \$6 billion a year by 2015.

"Our aim is to become the leading player in alternative energy," says John Browne, the BP veteran who has sought to cultivate the company's green image since he became chief executive in 1998. "We aim to grow the business five- or tenfold, and to establish it as a significant contributor to the restoration of energy security here in the United States and across the world," he told the Brookings Institution in Washington DC on 29 November 2005.

The strategy contrasts sharply with that

of BP's arch-rival, Texas-based ExxonMobil, whose fierce opposition to mandatory cuts in carbon emissions has riled environmental groups.

Yet not all those groups are fully convinced by BP's latest pronouncements on alternative energy. Rob Bradley, an energy specialist at the World Resources Institute, an environmental group in Washington DC, says that the investment is small compared with the company's massive annual profits. BP, he notes, is spending almost as much on advertising to build its green public image as it is ploughing into the new division. "They are never guilty of underselling what they are doing," he says.

The new division will be led by Steve Westwell, who previously ran BP's successful solar unit. One of its top priorities will be to find ways to cut carbon emissions from natural-gas power stations.

BP plans to road-test these technologies at a \$600-million plant in Peterhead, Scotland. Here, natural gas will be mixed with steam over a catalyst to produce hydrogen and, after another process, carbon dioxide. Burning the hydrogen will generate 350 megawatts of electricity, while the CO₂ will be pumped underground to flush out remaining oil from natural reservoirs under the North Sea.

So far, schemes such as this work only on paper. A decision about whether to move ahead

will be made at the end of 2006. Renewables specialists are watching with interest. "It is a complex new technology that has certainly not been proven at a commercial scale," says Doug Arent, a strategist at the US National Renewable Energy Laboratory in Golden, Colorado.

For these plants to be commercially viable, it would have to be cheaper for power companies to use them and sequester carbon underground rather than simply to emit carbon by burning raw natural gas. The US Department of Energy estimates that it costs \$150 a tonne to sequester carbon with current technologies; on Europe's young emissions trading markets, firms can buy the right to emit a tonne of carbon for about \$25 (see *Nature* **438**, 1077; 2005).

BP and other advocates of sequestration say that in time it will become cheaper to store carbon underground and more expensive to emit it. "I think they will make money ultimately, because I think the policy framework that rewards this will be developed," says Michael Liebreich, chief executive of New Energy Finance, a London consultancy that specializes in alternative energy sources.

BP's new division also incorporates its existing renewable-energy businesses, including the unit that makes solar panels. This holds about 10% of the fast-growing global market for such panels; last year, it generated revenues of more than \$400 million and turned in its first profit.

In addition, the company plans to erect and run more wind turbines. It already has two small wind farms in the Netherlands, and wants to build some larger ones in the United States, the first of which would generate about 200 megawatts of electricity.

The division also includes the oil company's combined-cycle gas-turbine business, which makes more efficient use of conventional gas turbines to provide heating for industrial or domestic use as well as to generate electricity. Not everyone would see that as alternative energy, but BP says it is just being practical.

"Our inclusion of it here is a pragmatic acceptance of the need for a mix of strategies," explains David Nicholas, a spokesman for BP.

"BP is never guilty of underselling what it is doing." — Rob Bradley

"We are trying to set this up as a business that will actually generate money and returns. It's kind of a hedge mix."

Markets reacted without any great enthusiasm to the launch of the unit: on the day of the announcement in late November, BP's stock slipped by 2.5%. And the scale of the move doesn't amount to much compared with the firm's 2004 profits of about \$15 billion.

"It's an investment," says Arent, "but they are not betting the company on it. I see it as the next step in their evolution from being an oil company to being an energy company." ■

Power games cause sparks in physics, but biologists have learnt from evolution

SIR — Ad Lagendijk, in his Essay “Pushing for power” (*Nature* 438, 429; 2005), draws a picture of the physics community dominated by aggressive males engaged in territorial combat and believes that these observations extend to other branches of science as well.

We beg to differ with him. We find biology to encompass a diversity of men and women, juniors and seniors, from around the globe.

At two conferences we have recently attended — the Ecological Society of America meeting in Portland, Oregon, in August 2004 and the International Botanical Congress in Vienna in July last year — women accounted for about a third of delegates and a similar proportion of session organizers and speakers.

Although women gave only about 19% of the keynote talks in Vienna, they were no less assertive than men in inviting themselves as keynoters for their sessions. (Sessions in Portland did not have keynote speakers, so cannot be compared.) Talks, whether good or bad, were rarely met by harsh or unfair criticism, and we did not feel that young scientists, non-native speakers or women were treated differently.

If physicists are to learn their lessons from evolution, they would do well to note that success does not come only to predatory males ‘red in tooth and claw’. It requires smart adaptation, soft power and, at more advanced levels, cooperation, all of which result in substantial diversity.

Perhaps they will also find that working in a more diverse community is more fun than being territorial predators.

Peter Hietz, Manuela Winkler

Department of Integrative Biology, University of Natural Resources and Applied Life Sciences, Gregor Mendel Strasse 33, 1180 Vienna, Austria

Power clashes limit science and reflect archaic values

SIR — The story told in Ad Lagendijk’s Essay “Pushing for power” (*Nature* 438, 429; 2005) is very familiar to many of us who have chosen to make a living in science.

As Lagendijk points out, the archaic and myopic value system now used in science favours a limited set of outcomes, often at the expense of scientific progress. But this value system and associated behaviours have other, far-reaching consequences.

First, they can compromise the integrity of the scientific enterprise

through tremendous pressure to publish.

Second, they may reduce the diversity of the workforce by attracting, rewarding and therefore retaining those who thrive in particular kinds of competitive environments — who are not necessarily those of greatest scientific ability, insight or creativity.

Third, they may limit the range and relevance of scientific pursuits and may devalue interactions between scientists and the public. Implicitly, this could reduce the potential to identify and address scientific issues of more global relevance.

The current reward system needs reassessing and reformulating in order to accommodate and reflect wider, more inclusive values.

María Uriarte*, Kathleen Weathers†, Valerie Eviner‡

*Columbia University, Department of Ecology, Evolution & Environmental Biology, 1200 Amsterdam Avenue, New York, New York 10027, USA

†Institute of Ecosystem Studies, Box AB, Millbrook, New York 12545, USA

Science is an adventure, not a battle

SIR — The world I inhabit, of experimental research in biological physics, is far from the cut-throat one described by Ad Lagendijk in his Essay “Pushing for power” (*Nature* 438, 429; 2005).

First, it is true that grants are often hard to obtain, yet almost everyone I know who has been willing to resubmit has ultimately been funded.

Second, many of the scientists I know — at least those who are involved in table-top experiments — have been able to push their own research agenda without being tempted or pressured to form egregiously large groups.

Third, although physics meetings can seem acrimonious, referee reports by physicists tend to be fair and matter-of-fact. I find this preferable to biology meetings, where talks may be greeted with “Let me compliment you on your wonderful data”, but new ideas are often rejected by grant and journal reviewers.

Fourth, in my experience many advisers shield their postdocs and allow them to make quiet progress.

There is certainly pressure to complete projects in a timely manner and for credit to be fairly parsed. But why should this be seen as an issue?

I do not wish to argue with Lagendijk, because we work in different fields in different countries. I simply want to let students and junior fellows realize that

doing science can be a lifelong adventure, rather than a constant battle.

David Kleinfeld

Department of Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0374, USA

Bias may be unintentional but it's still there

SIR — Marlene Zuk and Gunilla Rosenqvist, in their Correspondence “Evaluation bias hits women who aren’t twice as good” (*Nature* 438, 559; 2005), give an interesting view of women in science. Having been married to each other for almost two decades while raising two children and becoming professors at our respective universities in Taiwan, we would like to offer our perspective.

We agree that the under-representation of women in science, especially at higher levels, is primarily the result of bias in academic evaluation. The first author of this Correspondence was warned by her female adviser early in her career that women “have to try harder and do better than men, just to be able to compete with men”.

Actions intended to benefit women can also be used against them. A few years ago, the Taiwan legislature passed a workplace sex-equality law, in effect imposing quotas for women sitting on academic evaluation committees in Taiwan. One unhappy male faculty member questioned this quota by raising the question, in his university congress meeting: “Does a sex organ help a person to think [when evaluating]?”

Perceptions of male–female interactions can affect women in science too. At a faculty meeting determining the shortlist for hiring last year, Y.-H. H. spoke on behalf of the only eligible female candidate. Afterwards, he was asked by a junior faculty member whether the hiring of this young woman would “spell trouble” for his wife in the future.

Stavros Busenberg, a former professor of mathematics at Harvey Mudd College in California, who died in 1993, once commented on racism in the immediate aftermath of the 1992 Los Angeles riot: “In order not to be a racist, one must consciously try not to be one.” We would say this applies to discrimination of all types.

Cathy W. S. Chen*, Ying-Hen Hsieh†

*Department of Statistics, Feng Chia University, Taichung 407, Taiwan

†Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. They should be signed by no more than three authors; preferably by one.

BOOKS & ARTS

R. JUNO/CORBIS

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A robust approach

The functional overlap between different components protects biological systems.

Robustness and Evolvability in Living Systems

by Andreas Wagner

Princeton University Press: 2005. 408 pp.
\$49.50, £32.50

Eörs Szathmáry

When sitting on an aeroplane, we obviously hope that it won't crash. A tacit assumption behind this wish is that our biological system isn't about to crash either. It so happens that these systems share several features. Both have specific parts that serve certain functions. The plane was designed by engineers, who were in turn designed by evolution through natural selection. Both systems seem robust and yet fragile, but how can we reconcile these two seemingly opposing features? One answer is that they are robust and fragile to different perturbations, being particularly robust to perturbations that are common in their 'niche'. Another answer is that robustness can be in a trade-off with other features, such as price and reproduction rate.

In his book *Robustness and Evolvability*

in *Living Systems*, Andreas Wagner deals with some hot issues of current biology. As well as the terms in the title, the main keywords are 'neutral spaces', 'redundancy' and 'networks' — all of them highly fashionable, no doubt. But fashions fade, and usually a fraction of the original claims remain robust.

The robustness in biological systems is a consequence of their complexity. Many biologists are suspicious of ideas and models in the field of complexity, but I would encourage them to set their suspicions aside and read this book. Wagner's treatise is more than good biology; it is also very interesting biology. The picture is painted by talented hands. Wagner surveys many relevant examples, from the genetic code to organismal design, taking in properties of ribonucleic acids and proteins, the fascinating robustness of metabolic pathways and networks, the inelegant but robust genetic networks in development, and the many developmental pathways that can lead to essentially the same adult form. The level of detail is adequate in most cases and the questions and explanations are lucid. The

mathematical treatments are relatively easy to follow and deliver real insight.

One of the book's chief merits is that the author knows a lot about many levels of biological organization, something that can't often be said for the 'complexologists'. Many erroneous claims about, and faulty models of, various biological networks have sprung from a lack of knowledge about the complexity of biological phenomena.

If I have a favourite aspect of the book, it is the meticulous yet insightful analysis of neutral spaces and their relevance for the main themes of the book. Structures in biology can be envisaged as being embedded in a suitable 'space'. Protein-sequence space, for example, is multidimensional and discrete, as polypeptides are made up of a whole number of amino acids. There are 20 amino acids, so if the length of a polypeptide is, say, 100, then any given polypeptide chain has 1,900 (that is, 100×19) nearest 'neighbours', all of which differ by only one amino acid from the reference sequence. Many of these neighbours have the same biological function, because altering

one amino acid has little effect. But some proteins much farther away in this space (that is, with many different amino acids), may also have the same function. If you imagine that the space at all these 'neutral variants' is shaded, then the shaded areas make up the original protein's 'neutral space' and indicate the organism's robustness.

My only grumble is that I would have used the word 'domain', as the neutral region is a subregion of the whole protein space, but I hope no great confusion will arise. The author shows clearly that although neutral-space analysis supports old principles of population genetics in several cases, sometimes it does not. Consider, for example, the limited validity of the Haldane–Muller principle, which states

that the mutational load depends only on the mutation rate. Mean fitness is found to depend on the robustness of the population, which is actually a function of the structure of the neutral space.

I would like to know more about how the nervous system fits into the framework of Wagner's book. We know from studies ranging from the synapse to cell shape and functional neuroanatomy, for example, that at different scales a great many parts of the brain can have essentially the same function, in line with Wagner's analysis of 'distributed robustness'. But I realize that even the best story must end somewhere and leave space for future ones. ■
Eörs Szathmáry is at the Collegium Budapest, 2 Szentháromság utca, 1014 Budapest, Hungary.

Korea and subsequent work by the CIA, Stanley Milgram's obedience experiments and the Stanford prison experiment.

Much of Lemov's attention is given to the Yale Institute of Human Relations and the Rockefeller funding sources that supported it. Indeed, the extent to which Rockefeller money bankrolled laboratory psychology and allowed favoured schools of thought to flourish in the 1930s without a care for tomorrow, while most of the world was coping with the Great Depression, was news to me.

"A secret history that's not really a secret any more" is how the dustjacket puff characterizes Lemov's story. Unfortunately it's not really a history either. Her theme encompasses the whole of experimental, social and differential psychology, and more besides, but there is a great deal of selective focus in the tale she tells. The narrative veers from revealing new perspective to radical misconception, with some startling clangers on the way, and great chunks missing. A couple of howlers: on the use of electric shocks in experiments on rats, "the current ranged in intensity from 3.3 amperes to 7.6 volts"; on the use of a tape recorder in the 1950s, the tape was "running at the standard rate of seven and a half feet per second". There is also a bizarre misreading of history: rat researchers' accomplishments apparently "included the intelligence test, the SAT, the opinion survey, the early poll, the projective test..." (although the development of none of these is described at all). A huge omission, but for a passing reference, is operant conditioning and the work of B. F. Skinner. Such flaws undermine one's confidence in the author's sure-footedness on topics where one has to trust the detail of her account.

On a second reading, it becomes clear where the problem lies: Lemov has failed to get under the skin and into the minds of the characters who populate her narrative, and she has too selective and episodic a view of how psychology and related disciplines developed. It is as if she got her perspective from the sorts of stories that get into newspapers and failed to notice that they are but the tips of icebergs. Whether it is in the potted biographies or her account of laboratory methods, she comes to describe but stays only to scoff. The prevailing tone is one of uncritical pre-modern ethnography: look at this strange tribe and the weird things they got up to.

Empirical psychology arose and developed in contradistinction to prevailing views of the nature of mind, and grasped at emerging technologies both for theoretical models (hydraulic systems, the telephone exchange, the computer) and for laboratory tools. It is not necessary to agree with the theories, be sympathetic to the characters concerned or share their moral standpoint to make sense of what happened. Schools of thought flourished and then failed as their limitations were made manifest, but some left behind effective technical methods that can be used for good or ill, regardless of their theoretical origins.

Psychology in the real world?

World As Laboratory: Experiments with Mice, Mazes, and Men
by Rebecca Lemov

Hill & Wang: 2005. 304 pp. \$30

Steve Blinkhorn

The modern world was created by egg-headed, white-coated scientists working in laboratories, surrounded by complex equipment and inventing ever more ingenious gadgets. That at least is the popular myth of intentional or intelligent design, and there is some truth in it: after all, the laser began as a laboratory curiosity firmly rooted in hard science, but now delivers entertainment to the masses.

Other technologies, arising from theories long since discredited and from scientists whose claims to white-coat status now seem tenuous, have moulded our world more than we realize. What's more, there were shadowy and sometimes sinister sponsors at work promoting

and channelling developments that are now ubiquitous and inescapable in Western culture. These are techniques intended to shape behaviour, attitudes and thinking, arising principally from experimental psychology. And they are dangerous. That is the message I take from *World As Laboratory*, an anthropologist's view of twentieth-century psychological, behavioural and social science.

Rebecca Lemov's avowed focus is on the transfer of laboratory findings to the real world, and on the treatment of the real world as if it were some kind of laboratory. She finds manifest and manifold flaws in this enterprise, which she calls variously "human engineering" and "the American experiment". Her topics range from Jacques Loeb's experiments on tropisms, Elton Mayo's Hawthorne studies and Clark Hull's work on learning in rats, to the American administration of Pacific islands after the Second World War, brainwashing in

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

High hopes: in the 1950s researchers tried to manipulate behaviour by altering our view of the world.

It does not help an author attempting to communicate how this affected modern society to have an apparent aversion to technology and mathematics. On Hull's approach to learning theory, for example, she says, "from 1929 to 1936 the mechanisms become increasingly complex, like fine-spun webs". But she does not see that Hull's rather modestly complex work was a remote precursor of modern

attempts to create artificial intelligence, and she can scarcely have any acquaintance with the sorts of software models they involve. Hull just didn't have digital computers to play with. "Hull loved logarithms and seems to have had a fetish for mathematics," she remarks. Whether this would be a similarly perverse trait in a quantum physicist I cannot guess, but as most of the attempt to develop scientific

psychology has involved elements of quantification, I suppose her objection must be to that enterprise as a whole and in principle. It would be a more persuasive objection if it were based on a clearer, more comprehensive and more penetrating grasp of its target. ■

Steve Blinkhorn is at Psychometric Research and Development, Brewmaster House, The Maltings, St Albans AL1 3HT, UK.

EXHIBITION

Casting a long shadow

Melancholy: Genius and Insanity in the West

National Galleries of the Grand Palais in Paris, France, until 16 January 2006 and the Neue Nationalgalerie in Berlin, Germany, from 16 February to 7 May 2006.

Laura Spinney

In Albrecht Dürer's 1514 engraving *Melancholy*, a winged female creature sits in semi-darkness, surrounded by scientific symbols, resting her head on her hand, a pensive expression on her face. It is an image that has been reproduced again and again over the centuries, and it conveys the sense of solitude and depression of the spirit out of which springs both madness and the greatest acts of artistic and scientific invention. Dürer created his image in what might be called the 'golden age of melancholy', when that state of mind was closely linked with creativity.

The exhibition *Melancholy: Genius and Insanity in the West* sets out to track the changing attitudes to melancholy over the centuries. It shows how during different epochs it has been claimed by intellectuals, mystics, scientists and artists in turn. The ancient Greeks recognized the power of melancholic genius — what Aristotle called "the melancholy of exceptional men" — but with the arrival of Christianity this state of mind took on a diabolical significance.

The early Christians went out into the desert to wrestle with demons, which often appeared in the form of visions or obsessions. They were occasionally overcome by a kind of spiritual torpor that was regarded as a deadly sin. Medieval thinkers saw no creative energy in melancholy, and it was during the Middle Ages that the contemplative state was first associated with mania in the form of disturbing visions. This can be seen in Hieronymus Bosch's painting *The Temptation of St Anthony* and is reflected in our modern diagnosis of bipolar depression.

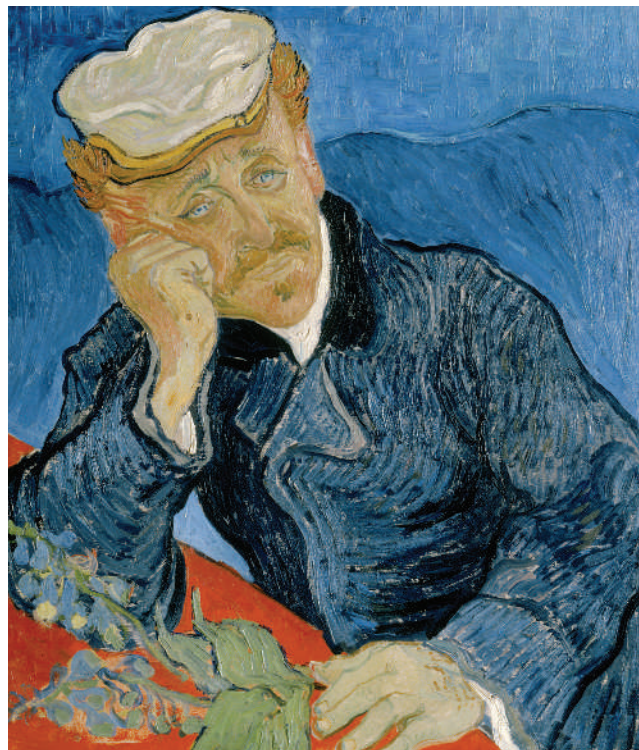
The Greeks attributed melancholy to the spleen and black bile. Autumn was its season, dusk its time of day, earth its element and Saturn its planet. After the Renaissance, the progressive Florentines resurrected melancholy as a

source of divine inspiration, and in Dante's *The Divine Comedy* it was from the sphere of Saturn that the poet saw the dazzling golden ladder that led upwards to contemplation of the deity.

The children of Saturn eventually came to include all social outcasts as well as people of melancholic temperament, and one of the themes in the exhibition is the loneliness of the monster. Goya's painting *Colossus* depicts a ghostly figure who strides over the landscape and the terrified, fleeing peasants; Charles Le Brun drew men resembling wolves; and the werewolf itself has a long association with melancholy. William Blake's Nebuchadnezzar from the painting of the same name is half man, half beast. Mad, naked and solitary, he crawls into a cave, fleeing the sun. The melancholic landscape is often bare and dotted with ruins. In the seventeenth century, Cervantes created the melancholic dreamer Don Quixote, but melancholy soon came to be seen as an affliction of island peoples, particularly the British.

Only music could reach the soul of the melancholic and lift his spirit, so David played his harp to soothe the disturbed King Saul. In Fernand Khnopff's painting *Listening to Schumann*, a woman whose face is hidden by her hand seems marooned as she sits in the centre of a drawing room, dressed in black, bathed in a sombre, late afternoon light. Schumann himself suffered from depression and eventually committed suicide.

With the arrival of the eighteenth century and the Enlightenment, the medical and artistic interpretations of melancholy diverged. In the world of science and rational thought, melancholy came to signify irrationality and madness. The medical diagnosis monomania was used to describe the obsessive aspect of



A picture of melancholy: Van Gogh's *Portrait of Dr Gachet*.

mania, and was later replaced by the broader diagnosis of bipolar depression, which covered both the manic and the depressive poles of the condition. To artists, in contrast, melancholy suggested solitude and meditation. Van Gogh's *Portrait of Dr Gachet* combines both interpretations. The artist's model studied melancholy at the Pitié-Salpêtrière hospital in Paris, and in the painting adopts the time-honoured pose of the wistful melancholic, head on hand. Meanwhile Ron Mueck's larger-than-life sculpture of a naked man cowering in a corner, chin in hand, speaks strongly of the asylum.

Science no longer has any use for the concept of melancholy, and the most widely accepted meaning now is the artistic one, what Victor Hugo described as "more than gravity and less than sadness". The modern concept of melancholy seems to be best summed up by Giorgio de Chirico's painting from 1912 showing a tomb bearing the marble figure of a reclining woman who rests her head on her hand. The tomb has engraved on it the word *melanconia*, and it casts a long, long shadow. ■

Laura Spinney is a freelance writer based in London and Paris.

NEWS & VIEWS

DEVELOPMENTAL NEUROSCIENCE

Two gradients are better than one

Liqun Luo

Wiring up retinal neurons to the correct brain region during development is a feat of precision growth. A novel directional cue repels retinal neuron fibres, acting as a counterbalance to a known attractive signal.

Our brain is made up of maps that organize what we sense. In the visual system, for example, an object is represented by the spatial activation pattern of retinal ganglion cells (RGCs), which form a two-dimensional sheet in the retina. RGC nerve fibres (axons) project into the brain in an orderly manner along both

the *x* and *y* axes, such that the two-dimensional image is recapitulated in the optic tectum region of the brain (Fig. 1). But how do maps like this form during development? Although the wiring diagram for how the RGC axons connect up with the tectum — the retinotopic map — has been extensively studied, it is still

not completely understood¹. On page 31 of this issue, Schmitt *et al.*² identify one of the signals that direct nerve fibres from the retina to their destination in the brain.

The retinotopic map was first elaborated by Roger Sperry 42 years ago³. By following point-to-point connections made as frog RGC axons regenerated between the retina and tectum, Sperry postulated that individual RGC axons must carry chemical tags that allow them to read the positional information in the tectum, also of a chemical nature. To limit the number of different tags needed to specify the connections, Sperry further proposed that the chemicals on RGC axons and in the tectum form gradients, such that the amounts of a tag could specify different positions. These ideas have been borne out spectacularly by experiment: first in the anterior–posterior axis, where gradients of a family of molecules called EphrinA in the tectum specify where RGC axons will end up^{4,5}; and more recently in the medial–lateral axis, where gradients of EphrinB molecules organize how the RGCs are wired up^{6,7}.

In the chick and the mouse, RGC axons home in on their exact targets along the medial–lateral axis in the tectum primarily by regulating the direction of branches that extend from the primary axons¹. When the primary axon ends up in a position lateral to where it should be, it projects branches medially to link up with its 'termination zone'; conversely, if the primary axon is medial to the termination zone, it branches out laterally (for example, the three purple neurons in Fig. 1). Graded expression of EphrinB molecules in the tectum and their receptors, the EphBs, in RGCs (Fig. 1; green gradients) regulate this direction of branching. RGCs that originate from ventral-most retina (Fig. 1; point 1) end up at the highest concentration of EphrinB (point A) because these RGCs have the most receptors, and therefore receive the most of the attractive EphrinB signal. In mutant mice that lack EphB2 and EphB3, individual axon branches preferentially extend laterally regardless of the position of primary axons relative to their termination zone, causing a lateral shift in RGC axon targeting⁸.

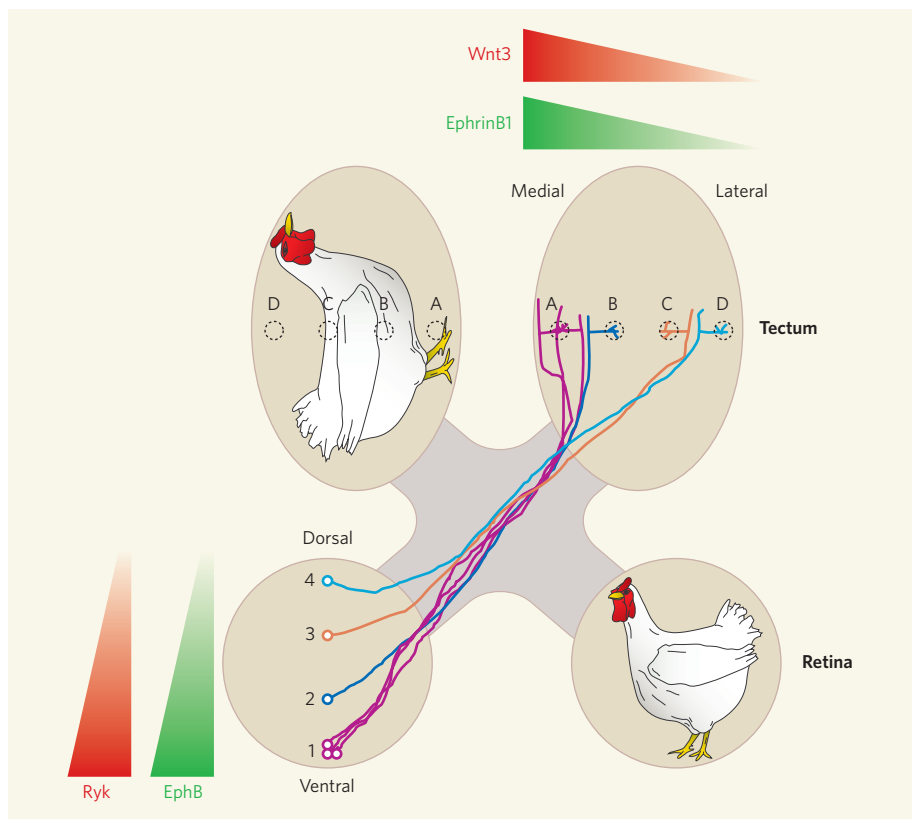


Figure 1 | Setting up the retinotopic map. An object on the retina (a chicken) is recapitulated in the tectum region of the brain. (The right retina is wired up to the left side of the brain and vice versa.) This is made possible by growth of nerve fibres (axons, coloured lines) from retinal ganglion cells (RGCs, coloured circles in the retina) along both *x* and *y* axes to the termination zones in the tectum (dotted circles). Only the retinal *y* axis, from dorsal to ventral, is depicted here. For instance, the purple neurons at the bottom perceive the bottom of the image, the chicken's feet; in the brain, they are wired up to region A on the medial side of the tectum, so the information from the image is transformed from the dorsal–ventral axis of the retina to a medial–lateral axis in the brain. The axons home in on their termination zones by regulating the direction of their branching — allowing the three purple axons to reach the same termination zone A. To achieve this, RGC axons expressing EphB receptors are attracted by the EphrinB gradient⁶. So, the purple axons with the most receptors are attracted to the medial side, where there is the most EphrinB. Schmitt *et al.*² have added a second gradient to this map, made of Wnt3 and its receptor Ryk. Repulsion of branches mediated by a Wnt3–Ryk gradient counterbalance the attraction mediated by EphrinB–EphB. (Adapted from refs 2 and 14.)

The directions provided by EphrinB/EphB alone, however, are not sufficient to account for the medial–lateral map. If they were, all RGC axons would head for the medial-most tectum, which is most attractive, and leave the lateral tectum unoccupied. Modelling studies suggest that an additional activity, most likely a repellent gradient in the same direction as the EphrinB attractive gradient, is necessary to counterbalance the medial-directing activity of EphrinB⁶. Schmitt *et al.*² now show that a gradient of the Wnt3 molecule is a strong candidate for this other directional signal.

Wnt3 belongs to the Wnt family of secreted proteins — classical regulators of development that specify a variety of cell fates in embryos using concentration gradients. Wnt proteins are implicated in many biological processes⁸, including axon guidance in fly embryos and in vertebrate spinal cord^{9–11}. Schmitt *et al.*² find that Wnt3 is expressed in the tectum in a gradient of the same direction as that of EphrinB. RGC axons express two different receptors for Wnt3, called Frizzled and Ryk. Frizzled receptors seem to promote axon-branch outgrowth, whereas Ryk inhibits it. Furthermore, Ryk is expressed in a gradient in RGCs in the same direction as the EphB receptors (Fig. 1; red gradients).

Schmitt *et al.* provide two lines of evidence to support the idea that Ryk-mediated repulsion in response to the Wnt3 gradient is partially responsible for RGC axon targeting along the medial–lateral axis. First, when Wnt3 is overexpressed in the tectum, RGC axons avoid the Wnt3 expression zone. Second, when Ryk activity is compromised in RGC axons, their termination zone in the tectum shifts medially. Notably, axon branches from Ryk-compromised RGCs selectively grow medially, regardless of their location in the tectum, a phenotype opposite to RGC axons that have mutant EphB. These data strongly suggest that Wnt3 repels RGC axons through the Ryk receptor. So, the Wnt3 concentration gradient provides a laterally directing force to counterbalance the medially directing force from EphrinB on individual RGC axon branches.

This study not only adds a missing piece to the retinotopic map, but also provides a satisfactory answer to the general question of how molecular gradients can direct the establishment of a neural circuit. As mentioned above, a single gradient–countergradient of ligand–receptor is not necessarily sufficient to specify an axis, and models propose that additional forces, such as a second counterbalancing gradient or axon–axon competition, are necessary. The identification of the Wnt3 gradient provides an example of how an attractive and a repulsive gradient in the same direction work together to specify an axis of a map.

One curious finding in the work by Schmitt *et al.*² is the direction of the Ryk receptor gradient in RGC axons. Previous modelling suggested that this second receptor could be

distributed in a gradient in the same direction as EphB, in the opposite direction, or without a gradient⁶. Given the observed direction of the gradient, and if EphrinBs and Wnt3 are the only two directional forces, then removing either EphrinBs or EphBs should cause an inversion of the map (inversion of 1→4 and A→D order in Fig. 1) in addition to a lateral shift. Mice that lack EphrinB or EphB are available, so it should already be feasible to test whether this is the case. Another prediction would be that the relative signalling strengths for the EphB and Ryk receptors differ for RGC axons originating from different positions along the dorsal–ventral axis in order to account for their differential behaviour in the tectum (imagine RGCs from points 1 and 2 in Fig. 1 arriving at the mid-point between A and B; to connect up properly, they have to branch out in opposite directions, even though their chemical environment is the same). More probably, additional directional forces may participate in these decisions, such as a potential attractive response to Wnt3 from the Frizzled receptors², a repulsive response to a high concentration of EphrinB¹², or activity-dependent refinement¹³. Future experiments

and modelling to assess the relative contribution of these forces and how they work together will surely enrich our understanding of map formation in the brain.

Liquan Luo is in the Howard Hughes Medical Institute, Department of Biological Sciences, Stanford University, Stanford, California 94305, USA.

e-mail: lluo@stanford.edu

- McLaughlin, T. & O'Leary, D. D. *Annu. Rev. Neurosci.* **28**, 327–355 (2005).
- Schmitt, A. M. *et al. Nature* **439**, 31–37 (2006).
- Sperry, R. W. *Proc. Natl Acad. Sci. USA* **50**, 703–710 (1963).
- Drescher, U. *et al. Cell* **82**, 369–370 (1995).
- Cheng, H.-J., Nakamoto, M., Bergemann, A. D. & Flanagan, J. G. *Cell* **82**, 371–381 (1995).
- Hindges, R., McLaughlin, T., Genoud, N., Henkemeyer, M. & O'Leary, D. D. *Neuron* **35**, 475–487 (2002).
- Mann, F., Ray, S., Harris, W. & Holt, C. *Neuron* **35**, 461–473 (2002).
- Logan, C. Y. & Nusse, R. *Annu. Rev. Cell Dev. Biol.* **20**, 781–810 (2004).
- Yoshikawa, S., McKinnon, R. D., Kokel, M. & Thomas, J. B. *Nature* **422**, 583–588 (2003).
- Lyuksyutova, A. I. *et al. Science* **302**, 1984–1988 (2003).
- Liu, Y. *et al. Nature Neurosci.* **8**, 1151–1159 (2005).
- McLaughlin, T., Hindges, R., Yates, P. A. & O'Leary, D. D. *Development* **130**, 2407–2418 (2003).
- McLaughlin, T., Torborg, C. L., Feller, M. B. & O'Leary, D. D. *Neuron* **40**, 1147–1160 (2003).
- Tessier-Lavigne, M. *Cell* **82**, 345–348 (1995).

PLANETARY SCIENCE

The ferryman casts his shadow

David J. Tholen

The most accurate way of determining the size of some bodies in the Solar System is to observe them as they pass across the face of a star. In the case of Charon, Pluto's largest satellite, it's been a long wait.

Ask people, especially children, to name their favourite planet, and often enough Pluto crops up. Whether that is due to the association with Disney's cartoon dog, or because of the enigmatic nature of the object — uniquely among the nine traditional planets, it has never been seen at close range by a spacecraft — isn't clear. What is certain, however, is that Pluto is an oddball. It doesn't quite fit the pattern of small, rocky planets such as Earth, or large, gassy ones such as Jupiter; indeed, over the past decade, attention has seemingly been riveted on the largely semantic question of whether Pluto is a planet at all.

This debate is unfortunate, as it has overshadowed some truly significant advances. On 31 October 2005, for example, it was announced that images taken by the Hubble Space Telescope had revealed two small satellites of Pluto, to add to its already familiar moon Charon. In this issue, Gulbis *et al.* (page 48)¹ and Sicardy *et al.* (page 52)² detail further progress in our understanding of the cold, distant Pluto system. They present observations of a stellar occultation — the passage of a Solar System object in front of a sufficiently bright

star — by Charon, allowing the most accurate assessment so far of the moon's size, and of the possibility that it has an atmosphere.

An occultation is a fairly rare, but extremely powerful, observational tool used by astronomers to make measurements of the occulting body that would otherwise be difficult or impossible to obtain. The gradual disappearance and reappearance of a star occulted by Pluto in 1988, for instance, showed that a thin atmosphere must envelop Pluto. It was 14 years before Pluto occulted a star again. When it did, it was found that its atmosphere had in the meantime approximately doubled in bulk — despite having moved farther away from the warming rays of the Sun. (Pluto has the least circular orbit of the traditional Solar System planets, its distance from the Sun ranging between 4.4 and 7.4 billion kilometres during an orbit lasting some 248 years.)

Pluto had been predicted to occult a star in 1980, but European observations confirmed that its shadow had missed Earth to the north. The recording of an occultation event at a South African observatory at the expected time therefore came as a surprise. The culprit

was Charon, whose orbit was not well known at the time (the satellite had been discovered only two years earlier), and which was fortuitously positioned south of Pluto, where its 750-mile-wide shadow would pass over the Earth. The South African observation was the only one made of that particular event, so, not knowing what part of the shadow had passed over the observation point, only a lower limit could be placed on the size of Charon³. Measurements of the star's brightness immediately before and after it disappeared behind Charon, however, left some astronomers wondering whether an extremely thin atmosphere had been detected, or whether the slight dimming they observed was simply due to low-probability statistical fluctuations.

Twenty-five years on, Gulbis *et al.*¹ and Sicardy *et al.*² present repeat measurements, for which they used some of the largest telescopes in the world situated high in the Andes mountains in Chile. The data place extremely low limits on the maximum density of an atmosphere; moreover, the combination of observations from telescopes at different locations on Earth's surface (and so at different positions under Charon's passing shadow; see Fig. 1 of refs 1 and 2) allowed an accurate assessment of the radius of curvature of the shadow. The new data therefore enabled both lower and upper limits to be placed on Charon's radius.

The constraints found on the radius (606 ± 8 km in one case¹, 603.6 ± 1.4 km in the other²) are consistent with each other. They can now be used to improve models of the system constructed around photometric observations of a series of occultation and eclipse events that occurred between 1985 and 1990, when Charon's orbit was seen approximately edge-on from Earth. Those observations were important in establishing the diameter of Pluto: the 1988 and 2002 stellar occultations by Pluto did not reveal its true size, because its thin atmosphere completely extinguished the starlight before it reached Pluto's surface. With the mass of the system established rather well from observations of the orbit of Charon around Pluto, the improved sizes will place tighter limits on the densities of both bodies; together with the discovery of two more satellites, this will further limit the mechanisms through which the system could have formed. As Pluto, seen from Earth, approaches the galactic plane with its much higher sky-density of stars, occultations should occur with increasing frequency, allowing further changes in Pluto's atmospheric bulk to be monitored, and perhaps a direct measurement to be made of the sizes of the two newly discovered satellites.

These exciting results^{1,2} come as the latest era in Pluto research is heralded by NASA's New Horizons mission. The month-long launch window opens on 11 January, and, if all goes according to plan, the spacecraft will fly past the Pluto–Charon system in July 2015. It's a frustratingly long wait for the scientific payoff, but there is little alternative for the Solar

System's more distant reaches. Coincidentally, some of the measurements to be made by the spacecraft will involve occultations by Pluto and Charon of the radio signals being beamed back to Earth.

David J. Tholen is at the Institute for Astronomy,

University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA.
e-mail: tholen@ifa.hawaii.edu

1. Gulbis, A. A. S. *et al.* *Nature* **439**, 48–51 (2006).
2. Sicardy, B. *et al.* *Nature* **439**, 52–54 (2006).
3. Walker, A. R. *Mon. Not. R. Astron. Soc.* **192**, 47–50 (1980).

OCEANOGRAPHY

A phosphate alternative

Sergio A. Sañudo-Wilhelmy

A major player among the phytoplankton can exploit a source of phosphorus previously thought to be unavailable to it. That ability may provide an ecological advantage in nutrient-depleted regions of the open ocean.

All marine phytoplankton depend on the availability of nutrients, and no nutrient is more important than nitrogen. Although there is an abundant supply of nitrogen in Earth's atmosphere, it is in the form of molecular N₂, which cannot be used by phytoplankton and must be 'fixed' before it can be utilized by most living organisms¹. The fixing process involves combining N₂ with hydrogen, and is carried out by organisms called diazotrophs. The most conspicuous marine diazotrophs are photosynthetic cyanobacteria of the genus *Trichodesmium*¹ — which, although they do not suffer from lack of nitrogen, require other essential elements such as iron and phosphorus for growth². Colonies of *Trichodesmium* thrive in vast open-ocean environments in which levels of readily available phosphorus (inorganic phosphate) — at some 10^{−9} mol per kg — are among the lowest in the world. So how do they manage to do so?

On page 68 of this issue³, Dyhrman *et al.* provide a plausible answer to this question. As well as using inorganic phosphate for growth⁴, *Trichodesmium* apparently mines the organic phosphonate compounds in the ocean's dissolved phosphorus pool. Many cyanobacteria,

including *Trichodesmium*, can hydrolyse other organic phosphorus compounds such as monophosphate esters (with C–O–P bonds)⁵, but it seemed that the tough carbon–phosphorus (C–P) bond in phosphonates⁶ put this source of phosphorus off limits to phytoplankton. Using a unique approach that combines genomic data with measurements of gene expression in laboratory cultures and natural assemblages of *Trichodesmium*, Dyhrman *et al.* have provided strong evidence to the contrary.

The authors performed a phylogenetic analysis using sequence data from the phosphonate (*phn*) genes that encode the enzymes essential to break C–P bonds. Their analysis showed that the C–P lyase enzymatic pathway required for phosphonate transport and hydrolysis is present in the genomes of all four of the *Trichodesmium* species sequenced to date, but is absent from the eight other available genomes of marine cyanobacteria. It also showed that the *phn* gene cluster appears in distantly related bacterial groups, implying that horizontal gene transfer has taken place (which is not unusual in bacterial genome evolution)⁷.

Similar phylogenetic results using the

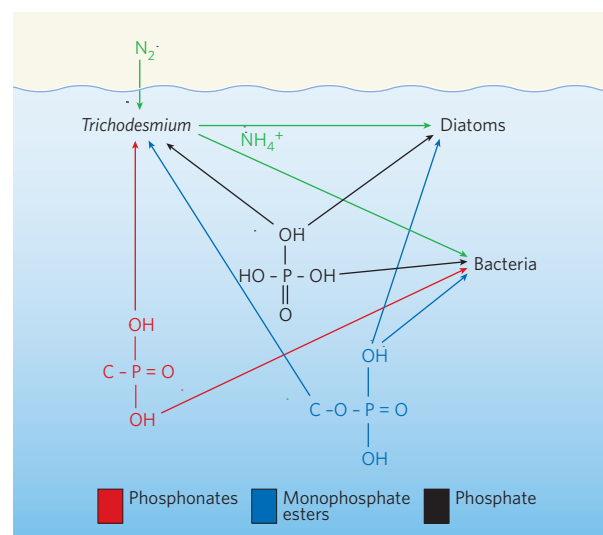


Figure 1 | Phosphorus compounds used by *Trichodesmium*.

Trichodesmium, a cyanobacterium, fixes atmospheric nitrogen, N₂, which it makes available to other marine organisms in the form of NH₄⁺. Dyhrman *et al.*³ now show that it can exploit phosphonates, organic compounds previously thought to be unavailable to phytoplankton, as well as monophosphate esters and inorganic phosphate. Bacteria can also use phosphonates. But other organisms such as diatoms, which are a major component of the phytoplankton, cannot.

sequences of the *phn* genes of the C–P lyase pathway have been reported⁸, but Dyhrman *et al.*³ have taken their investigations several steps further. The introduction of genetic material during horizontal gene transfer from one bacterial species to another does not guarantee that it will be functional in the recipient species. So Dyhrman *et al.* studied the expression of the phosphonate genes in *Trichodesmium* cultures grown under different conditions. Their results showed that genes of the C–P lyase complex are expressed only under conditions of phosphate starvation, similar to those found in nutrient-poor waters. Dyhrman *et al.* also looked for, and detected, *phn* gene expression in populations of *Trichodesmium* collected from the nutrient-depleted western North Atlantic.

It seems, then, that *Trichodesmium* has different strategies for exploiting a wide range of phosphorus compounds as nutrients, including phosphonates (Fig. 1). The limited information that exists suggests that phosphonates may be ubiquitous in the oceans⁹, making them a valuable phosphorus source for some phytoplankton. As dissolved organic phosphorus often constitutes the largest reservoir of this nutrient^{10,11}, drawing from this pool may give *Trichodesmium* an ecological advantage over other phytoplankton (including another oceanic diazotroph, *Crocosphaera watsonii*) that can apparently use only monophosphate esters or inorganic phosphate (Fig. 1). This may be an evolutionary adaptation to conditions in the nutrient-poor ocean gyres that are far from continental sources of nutrients. *Trichodesmium* species can form extensive blooms, up to some 300,000 km² in extent¹². Such blooms may in part be made possible by their unique ability to exploit a nutrient source that is not available to other nitrogen-fixing organisms.

As to remaining questions, it is still unclear what fraction of the phosphorus required by *Trichodesmium* is provided by phosphonates; even under extreme inorganic phosphorus limitation, phosphonates may be a relatively small part of the phosphorus budget. And it would be informative to include sequence data from a recently discovered unicellular nitrogen-fixing cyanobacterium¹³ in Dyhrman and colleagues' phylogenetic analysis. Because single cells may have lower nutrient requirements than do colonies such as *Trichodesmium*, it will be of interest to know whether or not these unicellular diazotrophs have the genetic capability for phosphonate uptake.

Further research on the cycling of phosphonates in marine systems is also called for, but that won't be an easy enterprise. Phosphonates were discovered in pre-concentrated high-molecular-weight organic matter in marine systems only eight years ago¹⁴. Although the presence of phosphonates and phosphorus esters can be determined using nuclear magnetic resonance spectroscopy^{9,14}, there is no available method for quantifying

phosphonates in sea water. The relative importance of this nutrient type under variable oceanographic conditions cannot be assessed without such measurements.

Finally, Dyhrman and colleagues' findings point to yet another role for *Trichodesmium* colonies in the open ocean. Just as the atmospheric nitrogen that they have fixed becomes available to other organisms, so those organisms may also benefit from the extra reactive phosphorus that *Trichodesmium* releases in excretory products or when it dies.

Sergio A. Sañudo-Wilhelmy is at the Marine Sciences Research Center, Stony Brook University, Stony Brook, New York 11794-5000, USA.
e-mail: ssanudo@notes.cc.sunysb.edu

1. Capone, D. G. *et al.* *Glob. Biogeochem. Cycles* **19**, GB2024; doi:10.1029/2004GB002331 (2005).
2. Mills, M. *et al.* *Nature* **429**, 292–294 (2004).
3. Dyhrman, S. T. *et al.* *Nature* **439**, 68–71 (2006).
4. Fu, F.-X., Zhang, Y., Leblanc, K., Sañudo-Wilhelmy, S. A. & Hutchins, D. A. *Limnol. Oceanogr.* **50**, 1459–1472 (2005).
5. Stihl, A., Sommer, U. & Post, A. F. *J. Phycol.* **62**, 310–317 (2001).
6. Hori, T., Horiguchi, M. & Hayashi, A. *Biochemistry of Natural C–P Compounds* (Maruzen, 1984).
7. Ochman, H., Lawrence, J. G. & Groisman, E. A. *Nature* **405**, 299–304 (2000).
8. Huang, J., Su, Z. & Xu, Y. *J. Mol. Evol.* **61**, 682–690 (2005).
9. Kolowitz, L. C., Ingall, E. D. & Benner, R. *Limnol. Oceanogr.* **46**, 309–320 (2001).
10. Ammerman, J. W. *et al.* *Eos* **84**, 165–170 (2003).
11. Karl, D. M. & Björkman, K. M. in *Biogeochemistry of Marine Dissolved Organic Matter* (eds Hansell, D. & Carlson, C.) 249–366 (Elsevier, 2002).
12. Dupouy, C. *et al.* *Eos* **81**, 13–16 (2000).
13. Zehr, J. *et al.* *Nature* **412**, 635–638 (2001).
14. Clark, L. L., Ingall, E. D. & Benner, R. *Nature* **393**, 426 (1998).

VOLCANOES

Interpreting inclusive evidence

Julia E. Hammer

Crystallization of ascending magma may affect the style of volcanic activity. Pockets of melt incorporated into crystals provide windows on processes that occur several kilometres below Earth's surface.

Most volcanic activity on land occurs above subduction zones, where one tectonic plate dives beneath another. When a subduction-zone volcano erupts, magma containing abundant dissolved H₂O ascends from a reservoir 8–14 km below the Earth's surface: it is the fate of H₂O vapour bubbles boiling out of the rising melt that largely determines whether the magma emerges with a bang or a whimper. Mount St Helens in Washington state displayed an impressive repertoire of eruptive styles in the early-to-mid 1980s. A magma intrusion produced a bulge on the volcano's flank in March 1980; eruption intensity peaked on 18 May with a sustained, explosive eruption of ash and pumice; and an uneasy denouement of brief explosions and lava extrusions characterized activity for the subsequent half-dozen years.

In a report in *Geology*, Blundy and Cashman¹ present new analyses of the products of Mount St Helens' activity during 1980. They interpret the compositions of frozen melt trapped within large crystals as snapshots captured through time during the magma's ascent, and ascribe variations in those compositions to sequential H₂O vaporization and crystallization of the melt. The authors suggest that crystallization occurs in response to decompression caused by degassing of the magma in isothermal conditions. They further propose that crystallization influences eruptive intensity through the resulting increase in magma viscosity.

The amount of H₂O that molten rock (a silicate melt) can hold depends strongly on pressure. If melt ascends from depth and decompresses, all H₂O in excess of the saturation

value enters a vapour phase. If the bubbles do not decouple from the melt and escape, their expansion causes magma to accelerate upwards. The result is further decompression, H₂O vaporization, magma expansion and so on, perpetuating a feedback loop that culminates in eruption. The importance of H₂O doesn't stop with dictating the physical properties of magma. Melting and freezing points of silicate melt are depressed by addition of H₂O, and this provides a thermodynamic driving force for solidification during magma ascent. For example, magmas at typical reservoir temperatures (about 850 °C) are only moderately crystalline when saturated with H₂O. If the melt dries out, extensive crystallization must occur to bring the system back into chemical equilibrium.

Plagioclase is a common crystalline mineral produced by magmas, and its stability is particularly sensitive to the amount of H₂O dissolved in the melt — the less H₂O, the more stable plagioclase becomes. Thus, plagioclase should crystallize in decompressing magmas. The importance of isothermal crystallization during volcanic eruptions is demonstrated by studies of volcanic materials^{2,3} as well as laboratory experiments^{4–6}. However, most work has focused on the formation of new crystals rather than the growth of existing ones⁷. In these studies, the usual assumptions have been that small plagioclase crystals (microlites) are created during an eruption⁸, whereas larger, compositionally heterogeneous crystals (phenocrysts) are inherited from the magma reservoir. In an earlier paper⁹, Blundy and Cashman questioned these assumptions: they proposed that

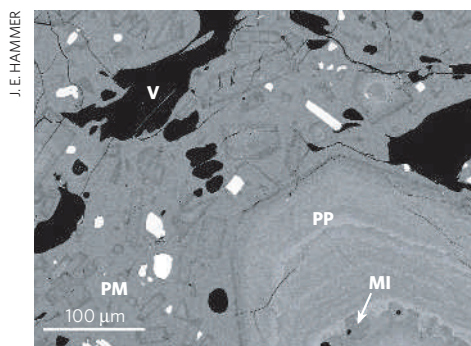


Figure 1 | Photomicrograph of a plagioclase phenocryst with a melt inclusion. This backscattered electron image shows the microtexture of a Mount St Helens lava-dome sample erupted in September 1981. PM, plagioclase microlite; PP, plagioclase phenocryst; MI, melt inclusion; V, vesicle (relict vapour bubble).

plagioclase phenocrysts form in a magma reservoir but then grow considerably during an eruption. In their new paper¹, they support and expand this idea with analyses of the melt inclusions found in phenocrysts (Fig. 1).

Melt inclusions are intriguing aberrations of 'normal' crystal growth. They preserve information about the history of the phenocryst, but are difficult to interpret because the conditions in which they form are unknown, the degree of chemical communication with the outside melt may vary, and post-entrapment processes can modify the original compositions. To address these problems, Blundy and Cashman¹ analysed more than 100 inclusions from six eruptions of Mount St Helens during 1980 that varied in intensity from quiescent lava effusion to sustained explosive activity.

They interpret a large range in observed H₂O contents (0.3–6.4 wt%) as representing saturation at pressures ranging from those at reservoir level all the way up to those near the surface. They also recognize a consistent relationship between the compositions of melt inclusions and the style of eruption that brought material to the surface. A key observation is that melt inclusions in lavas and other products of low-intensity eruptions preserve H₂O concentrations that correlate inversely with an 'incompatible species', potassium oxide (K₂O). Blundy and Cashman's finding that depressed H₂O is paired with elevated K₂O corroborates the notion that degassing and crystallization occur in concert. This trend among the melt inclusions suggests that connections to the surrounding melt persisted until the magma reached shallow levels; tube-like channels are in fact visible in some phenocryst cross-sections. What held these channels open initially, and what change in conditions triggered their closure, are grist for future studies.

In contrast, magmas that ascended rapidly during the explosive event of 18 May 1980 contain melt inclusions that have uniform K₂O yet variable H₂O. Blundy and Cashman interpret the few H₂O-rich inclusions in these magmas as snapshots of magma reservoir

conditions, and the remainder as the result of the eruption itself in which partial degassing of the melt happened too quickly for crystallization to occur. There is an alternative interpretation, however, that is consistent with the traditional understanding of plagioclase phenocrysts. Variable H₂O may arise from fluctuations in the CO₂ and H₂O content of magma within the reservoir. Such fluctuations would corroborate the view of the reservoir as an open system subject to periodic influxes of new magma¹⁰.

Most importantly, Blundy and Cashman's interpretation of the melt-inclusion data reinforces the idea that crystallization during an eruption affects the style of intermediate-intensity eruptions^{8,11}. The correlation between crystallinity and melt-inclusion H₂O content raises an intriguing chicken-and-egg issue, however. Does degassing-induced crystallization occur only when ascent rates and flow regimes in the magma conduit produce conditions favourable for rapid crystal growth? Or through its influence on magma viscosity, can the solidification process reduce the intensity of an eruption already in progress?

Events at Mount St Helens in 1980 have inspired many studies, the latest being this report¹ detailing changes in melt chemistry at unprecedented spatial and temporal resolution. With the volcano again obliging investigators with new magma since early 2004, a renewed effusion of research is sure to follow.

Julia E. Hammer is in the Department of Geology and Geophysics, University of Hawaii, 1680 East-West Road, Honolulu, Hawaii 96822, USA. e-mail: jhammer@soest.hawaii.edu

1. Blundy, J. & Cashman, K. *Geology* **33**, 793–796 (2005).
2. Swanson, S. E., Naney, M. T., Westrich, H. R. & Eichelberger, J. C. *Bull. Volcanol.* **51**, 161–176 (1989).
3. Cashman, K. V. *Contrib. Mineral. Petrol.* **109**, 431–449 (1992).
4. Swanson, S. E. *Am. Mineral.* **62**, 966–978 (1977).
5. Hammer, J. E. & Rutherford, M. J. *J. Geophys. Res.* **B107**, 8-1-8-24 (2002).
6. Couch, S. *Am. Mineral.* **88**, 1471–1485 (2003).
7. Geschwind, C. & Rutherford, M. J. *Bull. Volcanol.* **57**, 356–370 (1995).
8. Hammer, J. E., Cashman, K. V., Hoblitt, R. P. & Newman, S. *Bull. Volcanol.* **60**, 355–380 (1999).
9. Blundy, J. & Cashman, K. *Contrib. Mineral. Petrol.* **140**, 631–650 (2001).
10. Gardner, J. E., Carey, S., Rutherford, M. J. & Sigurdsson, H. *Contrib. Mineral. Petrol.* **119**, 224–238 (1995).
11. Gardner, C. A., Cashman, K. V. & Neal, C. A. *Bull. Volcanol.* **59**, 537–555 (1998).

BIOLOGICAL PHYSICS

Harmonies from noise

Michael Springer and Johan Paulsson

Do random environments make for random responses to them? Mathematical models suggest that this is not always the case — adding noise could create synchronous oscillations in cell–cell signalling systems.

Noise in communication devices is a familiar nuisance. In most Hollywood war films, radio static seems to botch up any attempt at coordinated action, to the frustration of the troops in the trenches. Cells face much the same problem: their signalling is garbled by chemical noise — random fluctuations in the concentrations of different molecular constituents — both inside and outside the cell. This noise could in turn compromise the cell's ability to grow and reproduce — or so one might think.

But two things here are worth considering more carefully. First, chemical fluctuations are always to some degree correlated, so different noise-afflicted cells may see the same random ups and downs. Second, in nonlinear systems (such as those underlying cell development, the cell cycle and circadian oscillators) the effects of the ups and downs do not cancel out; this in turn can qualitatively change the dynamics of the system. Writing in *Physical Review Letters*, Zhou *et al.*¹ propose a model for how the combination of these two effects can create regular and synchronized oscillations in an otherwise non-oscillatory cell system. This is an example of how noise in a biological process can have counterintuitive

effects, even suppressing other noise or generating new, coherent behaviours.

The investigations of Zhou *et al.* were inspired by a communication system between bacterial cells known as quorum sensing. Many bacteria produce a small 'autoinducer' molecule that, diffusing in and out of cells, promotes its own synthesis wherever it goes. This provides a population-wide positive-feedback loop that allows individual cells to count their neighbours and take synchronous action: when the population reaches a high enough concentration, the cells collectively switch from a low-production state, with minimal autoinduction, to a fully induced, high-production state.

The effect of changes in the design of quorum-sensing networks has been explored in several models. One proposal is to add a negative-feedback loop through an 'auto-inhibitor' molecule that, again diffusing in and out of cells, inhibits its synthesis wherever it goes. This additional loop creates a network similar to a circadian oscillator, in which concentrations go up and down in stable temporal waves; communication between cells by means of a diffusive autoinducer molecule could then allow these oscillations to be



50 YEARS AGO

The most important tercentenary of the year is undoubtedly that of the birth of Edmond Halley, the gifted mathematician, astronomer and pioneer meteorologist... In 1686 he made his famous visit to Newton at Cambridge which led the latter to prepare the "Principia". In 1686, after the first part had been delivered to the Royal Society, Halley not only undertook to supervise the printing, but also made himself responsible for the entire cost. This was because the finances of the Royal Society were so low that it was not only unable to carry out its original resolution to pay for printing the "Principia", but was even unable to pay Halley, then assistant secretary, his promised stipend of £50 a year. In July 1687, in the same month that Halley sent Newton twenty copies of the "Principia" to "bestow on your friends in the University", he accepted seventy copies of Willughby's "De Historia Piscium", published the previous year by the Royal Society, in lieu of some eighteen months salary. From *Nature* 7 January 1956.

100 YEARS AGO

The mounted skeleton of *Triceratops prorsus*... is interesting as displaying another Dinosaur of a distinct and very remarkable type. *Triceratops* was a quadrupedal reptile of quite moderate size, the skeleton... being not more than 25 feet in length and 10 feet in height... Two powerful horn-cores of the bovine type, 25 feet in length, rise from the frontal bones of the skull, at the base of which are the round bony orbits. The snout is narrow and pointed, and carries a third smaller horn upon the nasal bone. Behind the pair of frontal horns is an immense frill of bone spreading back over the occipital region and covering the first six vertebrae; it was 2 feet 6 inches long and 3 feet broad, resembling an immense Elizabethan ruff, ornamented with about twenty-four pointed bosses of bone along its border. From *Nature* 4 January 1906.

synchronized². Zhou *et al.*¹ analyse a simpler negative-feedback model, changing the natural autoinducer into an autoinhibitor that acts with some time delay. The system did not oscillate; instead, the level of autoinhibitor remained at a single steady state. But when enough random noise was added, stable and synchronized oscillations appeared.

How can this happen? Picture a mobile hanging from a baby's stroller. Although each pendulum of the mobile can swing back and forth periodically, none will actually move until a force is applied. The system is, however, poised to oscillate, and even a few petulant hits or a gust of wind can act as a trigger. If all the pendulums are affected at the same time and in the same way (in our analogy, if there is correlated extracellular noise), they can oscillate together. If the pendulums are connected by a string (if the cells can communicate), this is even easier to achieve: any pendulum that is out of synchrony is then literally pulled into phase by the others.

The analogy is not perfect: the combination of negative feedback and time delays can, under certain conditions, cause oscillations even without any external forces. And noise affects these conditions: even uncorrelated fluctuations can change the average concentration at which cell decisions are made, turn discrete all-or-nothing switches into smooth, continuous responses, or even sharpen continuous responses into discrete switches³. Uncorrelated noise could thus, in principle, sufficiently change the characteristics of a non-oscillating feedback system to produce stable oscillations^{3,4}.

This goes against our expectation that noise blunts a signal. In fact, the combination of noise and nonlinear kinetics is capable of almost anything, even in the simplest systems. To take an example from biochemistry, the rate at which two identical protein monomers dimerize to form an active enzyme depends on the square of the monomer concentration. Cells with higher than average monomer levels thus contribute disproportionately to the average rate of dimerization in the population. If each cell had exactly one monomer, no enzyme would be produced. But if fluctuations were such that half of the cells had two monomers and the other half none, some enzyme would be made even though the average amount of monomer is the same⁵. Because protein fluctuations in turn respond very differently to the rates at which the genes encoding them are transcribed and translated⁶, changing both rates simultaneously can have unexpected effects on the average rate of enzyme production, potentially even making it proportional to the cube, rather than the square, of the average monomer concentration.

Correlated fluctuations between different components in the same cell can be even more useful. Studies using two identically regulated alleles (gene variants) that encode fluorescent 'reporter' proteins show that some protein

noise is correlated and shared between alleles, but that other sources of protein noise are uncorrelated and experienced separately⁷. From the viewpoint of each gene individually, there is little difference between the two types of noise; crucially, however, shared noise can be used to coordinate action between them. To continue the biochemical example, consider this time two different types of monomer that combine to form an active enzyme complex: if the noise in the expression of the two monomers were correlated, randomly pushing up both concentrations at the same time, cells could form enzyme complexes more efficiently.

Zhou and colleagues' work¹ is just one example of the growing body of research that shows how cells can use noise to suppress other noise, or to create oscillations, multi-stabilities and many other coherent kinetic traits. Coupled with the recognition that even simple, noise-free mechanisms can generate 'deterministic chaos' — acting to amplify infinitesimal environmental perturbations into large variations in physiological characteristics⁸ — this line of research subverts the simplest picture of noise. That is that cells exploit noise when they need heterogeneity, and suppress it when they need deterministic, coherent behaviours.

This sea change in our perception raises an important and almost entirely unaddressed question. We know that cells can create virtually any type of noise; we also know that they can create almost any type of nonlinearity. What we rarely know, however, is which of the available strategies cells actually use. Is there a grander scheme that explains the choice between deterministic and noise-driven solutions? The only way to address this question experimentally is to characterize the single-cell dynamics of a large number of systems, and to see which strategies tend to be used to solve which problems. Given some reasonable physical constraints — mechanistic, energetic, evolutionary or otherwise — it may even be possible to partially answer the question from first principles. ■

Michael Springer and Johan Paulsson are in the Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02108, USA. Johan Paulsson is also in the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK. e-mail: johan_paulsson@hms.harvard.edu

1. Zhou, T., Chen, L. & Aihara, K. *Phys. Rev. Lett.* **95**, 178103 (2005).
2. McMillen, D., Kopell, N., Hasty, J. & Collins, J. J. *Proc. Natl Acad. Sci. USA* **99**, 679–684 (2002).
3. Horsthemke, W. & Lefever, R. *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology* (Springer, Berlin, 1984).
4. Vilar, J. M., Kueh, H. Y., Barkai, N. & Leibler, S. *Proc. Natl Acad. Sci. USA* **99**, 5988–5992 (2002).
5. Raleigh, E. A. & Kleckner, N. *Proc. Natl Acad. Sci. USA* **83**, 1787–1791 (1986).
6. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. *Nature Genet.* **31**, 69–73 (2002).
7. Elowitz, M. B. *et al. Science* **297**, 1183–1186 (2002).
8. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Perseus, Cambridge, MA, 1994).

BRIEF COMMUNICATIONS

Deep-sea fishes qualify as endangered

A shift from shelf fisheries to the deep sea is exhausting late-maturing species that recover only slowly.

Criteria from the World Conservation Union¹ (IUCN) have been used to classify marine fish species as endangered since 1996, but deep-sea fish have not so far been evaluated — despite their vulnerability to aggressive deepwater fishing as a result of certain life-history traits². Here we use research-survey data to show that five species of deep-sea fish have declined over a 17-year period in the Canadian waters of the northwest Atlantic to such an extent that they meet the IUCN criteria for being critically endangered. Our results indicate that urgent action is needed for the sustainable management of deep-sea fisheries.

At one time it was presumed from the vastness of the oceans that fishing would not drive species to extinction. There have, however, been recent sharp declines in the numbers of oceanic cod, sharks, rays, tuna, marlins, swordfish and sea turtles^{3–7}. As the shelf fisheries in the northwest Atlantic began to collapse in the 1960s and 1970s, harvesting shifted to deep-sea fish species⁸, but many populations crashed within ten years because their recovery is so slow^{2,9}.

Deep-sea fish are highly vulnerable to disturbance because of their late maturation, extreme longevity, low fecundity and slow growth^{2,9}. Some deep-sea fish form spawning aggregations on seamounts and the sea floor, and this increases their susceptibility to overfishing². Survey data collected over extended periods are limited so it has been difficult to determine the effects of deep-sea fishing on both target and by-catch species.

For our analysis, we chose five species that live on or near the bottom of the North Atlantic Ocean, on the continental slope. They range from the common to the rare: roundnose grenadier, *Coryphaenoides rupestris*; onion-eye grenadier, *Macrourus berglax*; blue hake, *Antimora rostrata*; spiny eel, *Notacanthus chemnitzii*; and spinytail skate, *Bathyraja spinicauda*. The species evaluated can live to 60 years of age, grow to more than 1 m in length, and mature in their late teens. *C. rupestris* and *M. berglax* have been commercially fished, and all five are taken as by-catch in fisheries that target Greenland halibut, *Reinhardtius hippoglossoides*, and redfish, *Sebastes* spp. None was taken in any substantial number, even as by-catch, before the 1970s.

We used catch data from standardized, research-trawl surveys in the Canadian waters of the northwest Atlantic Ocean over 1978–94

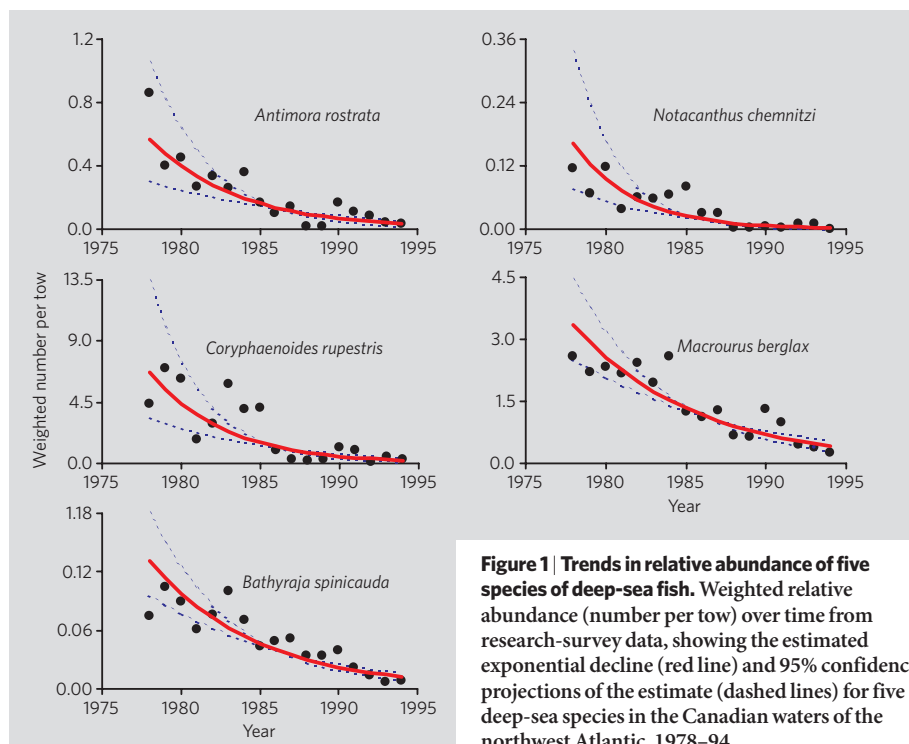


Figure 1 | Trends in relative abundance of five species of deep-sea fish. Weighted relative abundance (number per tow) over time from research-survey data, showing the estimated exponential decline (red line) and 95% confidence projections of the estimate (dashed lines) for five deep-sea species in the Canadian waters of the northwest Atlantic, 1978–94.

to determine declines in relative abundance and individual mean size (for details, see supplementary information). All species declined in relative abundance (Fig. 1): declines over the 17-year period were 87–98% and declines estimated for three generations, the IUCN benchmark, were 99–100% (see supplementary information). Survey data for an additional period (1995–2003) were obtained for *C. rupestris* and *M. berglax*. The overall declines in relative abundance for these two species over the 26-year period were 99.6% and 93.3%, respectively; estimated declines over three generations were 100% and 99.7%, respectively (see supplementary information).

According to the IUCN criteria, these five species of deep-sea fish qualify as critically endangered in the northwest Atlantic. The declines occurred on a timescale equal to, or slightly less than, a single generation of these species. All of the species apart from *N. chemnitzii* also declined by 25–57% in mean size over the 17-year period (see supplementary information). The survey data are not adequate for full assessment of the situation for other deep-sea fish species that may also be at risk. The largest deepwater skate in the northwest Atlantic — the barndoor skate *Dipturus laevis*

— was driven unnoticed almost to extinction⁶.

Scientific investigation lags behind the collapse of deep-sea fisheries^{8,9}. Conservation measures are necessary and lack of knowledge must not delay appropriate initiatives, including the establishment of deep-sea protected areas.

Jennifer A. Devine, Krista D. Baker, Richard L. Haedrich

Department of Biology, Memorial University, St John's, Newfoundland A1C 5S7, Canada
e-mail: jdevine@mun.ca

1. World Conservation Union IUCN Red List Categories and Criteria: Version 3.1 (2001).
2. Koslow, J. A. et al. *ICES J. Mar. Sci.* **57**, 548–557 (2000).
3. Pauly, D. et al. *Nature* **418**, 689–695 (2002).
4. Myers, R. A. & Worm, B. *Nature* **423**, 280–283 (2003).
5. Baum, J. K. et al. *Science* **299**, 389–392 (2003).
6. Casey, J. M. & Myers, R. A. *Science* **281**, 690–692 (1998).
7. Graham, K. J., Andrew, N. L. & Hodgson, K. E. *Mar. Freshwat. Res.* **52**, 549–561 (2001).
8. Haedrich, R. L., Merrett, N. R. & O'Dea, N. R. *Fish. Res.* **51**, 113–122 (2001).
9. Moore, J. A. & Mace, P. M. *Fisheries* **24**, 22–23 (1999).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/439029a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

ECOLOGY

Mechanisms for consumer diversity

Arising from: W. A. Nelson, E. McCauley & F. J. Wrona *Nature* 433, 413–417 (2005)

A variety of mechanisms can theoretically produce competitive coexistence in nature¹, making it hard to identify a single explanation for the maintenance of diversity in any particular system. Based on laboratory experiments with a consumer–resource system of crustacean *Daphnia* eating algae, Nelson *et al.*² suggest that maintenance of genetic diversity in the consumer populations they studied depends only on the dynamics of the population structure of the consumer. We suggest that the differences in *Daphnia* genetic diversity that they find for different experimental treatments could equally be explained by a simple, well known mechanism: the number of coexisting competitors cannot exceed the number of shared resources^{3–5}. Here we confirm this possibility by using a simple mathematical model and suggest that more than one mechanism may account for the maintenance of genetic diversity observed by Nelson *et al.* in their system.

Nelson *et al.*² study the response of *Daphnia* molecular genetic diversity to differently driven population cycles of a *Daphnia*–algae system. Experimental treatments consisted of “driven experiments” that produced predator–prey cycles by forcing the system externally, and “coupled experiments,” in which the system had internally driven stage-structured cycles. Both types of dynamic have been described previously⁶.

One of several potentially confounding factors in the treatments employed by Nelson *et al.* is the different number of algal species used. The observed dynamics of *Daphnia* population structure were strikingly different between the two treatments. The authors propose that the differences observed in *Daphnia* genetic diversity between treatments were the result of the presence or absence of internally generated, stage-structured population cycles. Although their explanation is plausible, other explanations should be tested by mathematical modelling.

We propose an alternative explanation for the differences in *Daphnia* genetic diversity. A single green algal species was provided in the driven experiments, in which Nelson *et al.* observed rapid competitive exclusion of all but a single *Daphnia* genotype. By contrast, four algal species were provided in the coupled experiments, in which the authors observed reduced magnitude of selection and consequently claimed that genetic diversity was maintained. Reducing this scenario to the simplest possible case, we built a model with two zooplankton genotypes (without age structure) and two algal species growing in a

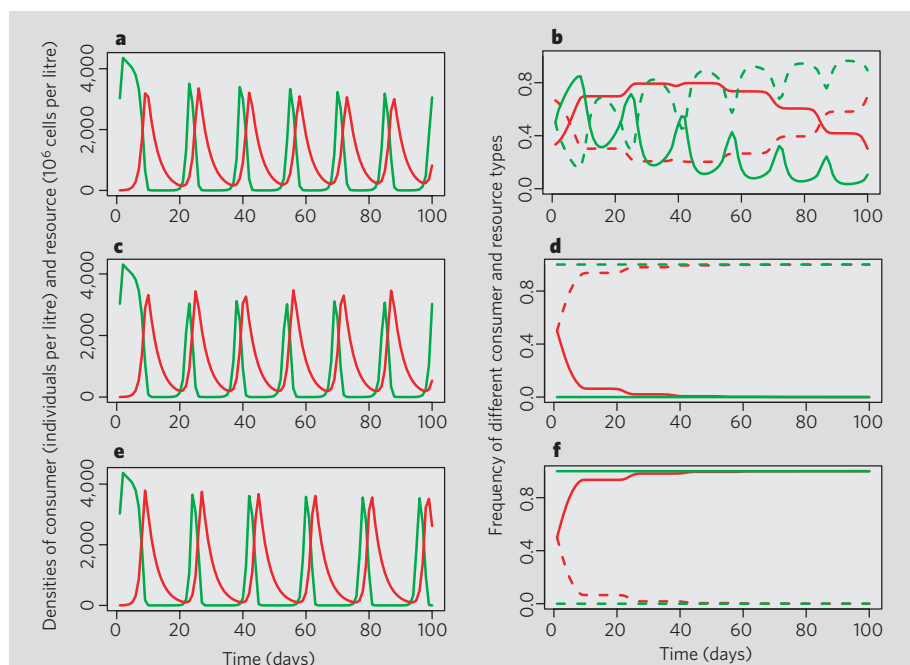


Figure 1 | Effects of resource diversity on maintenance of consumer genetic diversity. Simulations are run at a dilution rate $\delta = 0.3$ per day. Dashed lines: consumer 1 (red), which benefits most from resource 1 (green); solid lines: consumer 2 (red), which benefits most from resource 2 (green). **a, b**, Two consumer types and two resource types are inoculated: all persist and community diversity is maintained. **c, d**, Both consumers are inoculated with only resource type 1: consumer 1 persists and consumer 2 dies out. **e, f**, Both consumers are inoculated with only resource type 2: consumer 1 becomes extinct and consumer 2 persists. Methods: we extended the model of Yoshida *et al.*¹⁰ to include two grazers R_1, R_2 feeding on clonal algal populations C_1, C_2 in a chemostat that had one nutrient limiting algal growth. Algal clone i is characterized by its vulnerability to predation, p_i , and by ‘food qualities’, q_{ij} , representing its value to consumer j . Decreased vulnerability to predation has a cost in terms of the half-saturation constant for nutrient uptake. In these simulations, $p_1 = 0.75$ and $p_2 = 1$; food qualities are $q_{1,1} = q_{2,2} = 1$ and $q_{1,2} = q_{2,1} = 0.75$. Other parameter values and definitions are as listed in supplementary Table 1 of Yoshida *et al.*¹⁰, but with the following changes: $\alpha_1 = 1$, $\alpha_2 = 40.0$, $\chi_B = 3,400$.

flow-through (chemostat) system in which a single nutrient limited algal growth.

Zooplankton are generalist consumers, but each type benefits most (per algal cell consumed) from a different algal species. Typically, the result is slow selection or coexistence of zooplankton genotypes when the system has both algal species present, and fast selection, proceeding to fixation, when it is inoculated with either of the two algal species alone (Fig. 1). Our model results are strikingly similar to the data of Nelson *et al.*². Two different food resources are considered to be prerequisite for the coexistence of two competing species⁷ and coexistence of zooplankton competitors is possible if two different food resources are given in appropriate proportions^{8,9}.

We suggest that the mechanisms underlying the experimental results of Nelson *et al.*² need further investigation before we can accept their conclusion that stage-structured

dynamics dominate in the maintenance of genetic diversity. Mathematical modelling or additional supporting data from the reported experiments could be used to rule out other possibilities; further experiments could eliminate potentially confounding differences between the two experimental treatments.

Takehito Yoshida, Laura E. Jones,

Stephen P. Ellner, Nelson G. Hairston Jr

Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853-2701, USA

e-mail: ty59@cornell.edu

- Chesson, P. *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).
- Nelson, W. A., McCauley, E. & Wrona, F. J. *Nature* **433**, 413–417 (2005).
- Volterra, V. J. *Cons. Int. Explor. Mer* **3**, 3–51 (1928).
- Hardin, G. *Science* **131**, 1292–1298 (1960).
- MacArthur, R. & Levins, R. *Proc. Natl Acad. Sci. USA* **51**, 1207–1210 (1964).
- McCauley, E., Nisbet, R. M., Murdoch, W. W., de Roos, A. M. & Gurney, W. S. C. *Nature* **402**, 653–656 (1999).

7. Tilman, D. *Resource Competition and Community Structure* (Princeton Univ. Press, Princeton, 1982).
8. Rothhaupt, K. O. *Nature* **333**, 660–662 (1988).
9. Ciroso-Perez, J., Carmona, M. J. & Serra, M. *Limnol. Oceanogr.* **46**, 1511–1523 (2001).

10. Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F. & Hairston, N. G. Jr *Nature* **424**, 303–306 (2003).

doi:10.1038/nature04526

ECOLOGY

Nelson et al. reply

Replying to: T. Yoshida, L. E. Jones, S. P. Ellner & N. G. Hairston Jr
Nature doi:10.1038/nature04526 (2005).

We have demonstrated that qualitatively different consumer–resource dynamics can have a large impact on natural selection in a consumer population, and proposed that the change in selection is generated by the juvenile–adult stage–structure of the consumer¹. Yoshida *et al.*² propose an alternative mechanism to explain our results, but, on the basis of other evidence collected at the time of our experiments¹, we can refute this explanation. We also discuss theoretical results that show how simple stage–structure in a consumer population can modify selection.

The consumer–resource interaction between *Daphnia* and algae shows two qualitatively different classes of population cycle: small-amplitude, stage-structured cycles and large-amplitude, consumer–resource cycles³. These cycles are characterized by different amplitudes, frequencies and juvenile development rates. Among competing genotypes of *Daphnia*¹, we found that small-amplitude, stage-structured cycles reduce the amount of selection compared with large-amplitude, consumer–resource cycles — a result that we explained in terms of the interaction between resource dynamics and resource-dependent juvenile development, which emerges from general competition models based on *Daphnia* and algal ecology^{4–7}.

Yoshida *et al.*² propose that the observed change in selection results from the difference in algal diversity between treatments, rather than a difference in population dynamics. They suggest that the four algal species provided in the stage-structured treatment allows for trade-offs among genotypes, whereas the single algal species in the consumer–resource treatment does not. This mechanism would require that *Daphnia* genotypes have different performance (relative to one another) with each algal species, and that selection is slow when multiple algal species are present and rapid when only one is present.

However, these features do not exist in our experiments. The algae we used are easily ingested by *Daphnia* and, as *Daphnia* are gen-

eralist filter-feeders⁸, differential ingestion of algal species is unlikely. Furthermore, *Daphnia* genotypes do not show the required trade-off in performance among algal species⁹. A more direct test is to compare the rate of selection among genotypes against the dynamics of each algal species. In our experiments, the algal community became dominated by a single species at around day 50 (Fig. 1a) — a common observation in *Daphnia*–algal aquaria experiments¹⁰. As this does not lead to the required rapid competitive exclusion (Fig. 1b), the mechanism suggested by Yoshida *et al.* can be rejected.

We agree that it is challenging to identify the mechanisms that maintain diversity, particularly when they involve life-history features that are difficult to manipulate. However, the alternative mechanism proposed by Yoshida *et al.*² cannot capture the main result of our experiments — selection is reduced under small-amplitude population cycles¹. Theoretical work on physiologically structured models in *Daphnia* has tightly linked the models to experiments^{5–7}. Based on those results, we have developed a competition model to test the influence of structured *Daphnia*–algae population dynamics on selection among *Daphnia* genotypes⁴. This independently parameterized model predicts both of the two classes of population cycle. Selection among competing consumer genotypes changes depending on the class of dynamics — selection under large-amplitude cycles is greater than under small-amplitude cycles.

As we proposed¹, the difference in selection is caused by the influence of resource dynamics on mortality and juvenile development. Our experimental and theoretical results both indicate that structured population dynamics can have a strong influence on natural selection through a potentially general equalizing mechanism.

William A. Nelson[†], Edward McCauley*, Frederick J. Wrona‡

*Ecology Division, University of Calgary, Calgary, Alberta T2N 1N4, Canada

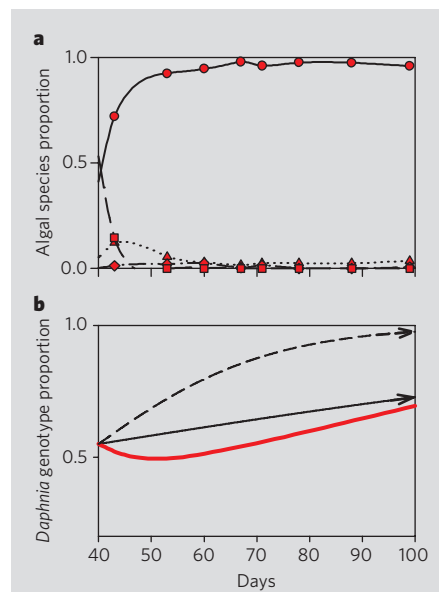


Figure 1 | Algal composition and genotype selection. **a**, Example of species composition of the algal community changing with time (see also Fig. 2a in ref. 1). Lines represent individual algal species and symbols indicate times at which samples were taken. **b**, Selection dynamics for genotype '9H'. Red line shows selection for example shown in **a**; black line shows average selection under stage-structured cycles; and dashed line shows the faster selection under consumer–resource cycles. The genotype dynamics in stage-structured cycles do not show rapid competitive exclusion when the algal population is dominated by a single species.

†Present address: Centre for Mathematical Biology, University of Alberta, Edmonton, Alberta T6G 1G1, Canada

e-mail: wanelson@math.ualberta.ca

‡Water & Climate Impacts Research Centre, National Water Research Institute, University of Victoria, Victoria, British Columbia V8W 3P5, Canada

1. Nelson, W. A., McCauley, E. & Wrona, F. J. *Nature* **433**, 413–417 (2005).
2. Yoshida, T., Jones, L. E., Ellner, S. P. & Hairston, N. G. Jr *Nature* **438**, doi:10.1038/nature04526 (2005).
3. McCauley, E., Nisbet, R. M., Murdoch, W. W., de Roos, A. M. & Gurney, W. S. C. *Nature* **402**, 653–656 (1999).
4. Nelson, W. A., McCauley, E. & Nisbet, R. M. *Ecol. Lett.* (submitted).
5. McCauley, E., Nisbet, R. M., de Roos, A. M., Murdoch, W. W. & Gurney, W. S. C. *Ecol. Monogr.* **66**, 479–501 (1996).
6. Noonburg, E. G. *et al. Funct. Ecol.* **12**, 211–222 (1998).
7. Nisbet, R. M., McCauley, E., Gurney, W. S. C., Murdoch, W. W. & Wood, S. N. *Ecology* **85**, 3132–3139 (2004).
8. DeMott, W. R. *Limnol. Oceanogr.* **27**, 518–527 (1982).
9. Repka, S. *Freshwat. Biol.* **38**, 675–683 (1997).
10. Nelson, W. A., McCauley, E. & Wrona, F. J. *Proc. R. Soc. Lond. B* **268**, 1223–1230 (2001).

doi:10.1038/nature04527

Wnt-Ryk signalling mediates medial-lateral retinotectal topographic mapping

Adam M. Schmitt^{1*}, Jun Shi^{1*}, Alex M. Wolf², Chin-Chun Lu¹, Leslie A. King² & Yimin Zou^{1,2,3}

Computational modelling has suggested that at least two counteracting forces are required for establishing topographic maps. Ephrin-family proteins are required for both anterior-posterior and medial-lateral topographic mapping, but the opposing forces have not been well characterized. Wnt-family proteins are recently discovered axon guidance cues. We find that *Wnt3* is expressed in a medial-lateral decreasing gradient in chick optic tectum and mouse superior colliculus. Retinal ganglion cell (RGC) axons from different dorsal-ventral positions showed graded and biphasic response to *Wnt3* in a concentration-dependent manner. *Wnt3* repulsion is mediated by Ryk, expressed in a ventral-to-dorsal decreasing gradient, whereas attraction of dorsal axons at lower *Wnt3* concentrations is mediated by Frizzled(s). Overexpression of *Wnt3* in the lateral tectum repelled the termination zones of dorsal RGC axons *in vivo*. Expression of a dominant-negative Ryk in dorsal RGC axons caused a medial shift of the termination zones, promoting medially directed interstitial branches and eliminating laterally directed branches. Therefore, a classical morphogen, *Wnt3*, acting as an axon guidance molecule, plays a role in retinotectal mapping along the medial-lateral axis, counterbalancing the medial-directed EphrinB1-EphB activity.

Topographic mapping of axonal connections is a fundamental and widespread feature of nervous system wiring, whereby spatial orders of neurons are smoothly and continuously mapped to their targets. Molecular labels in the target fields specify topographic connections by activating receptors expressed in growth cones. Classical studies and computational modelling propose that balanced opposing forces are necessary for generating smooth topographic maps^{1–4}. The ephrin family of proteins have been identified as axon guidance cues important for map formation.

The retinotectal projections are organized along the anterior-posterior and dorsal-ventral axes. Temporal axons terminate at anterior tectum, and nasal axons project to the posterior tectum. Ventral retinal axons connect to medial (dorsal) tectum, and dorsal retinal axons find their targets at lateral (ventral) tectum. Studies implicated A-class ephrins in establishing anterior-posterior topographic mapping via a repulsive mechanism through the EphA receptors^{5–8}. Along the medial-lateral axis, an attractive interaction involving EphrinB–EphB was proposed to control dorsal-ventral patterning^{9,10}.

Our previous work identified Wnts as guidance molecules for ascending spinal cord commissural axons¹¹ and descending corticospinal tract axons¹⁶. Here we report evidence that a classical morphogen, *Wnt3*, acts as an axon guidance molecule and plays an essential role in medial-lateral topographic map formation by counterbalancing the opposing medial-directed cue, EphrinB1¹⁰. The identification of *Wnt3* as a counterbalancing cue to EphrinB1 provides the first direct experimental evidence for theoretical models of counterbalancing forces in topographic neural map formation^{1,2}.

Wnt3 is expressed in a medial-lateral decreasing gradient

In situ hybridization showed that *Wnt3* is expressed in a high-to-low gradient along the medial (dorsal)-to-lateral (ventral) axis in the ventricular zone in chick optic tectum at embryonic day 10 (E10) (Fig. 1a) and mouse superior colliculus at postnatal day 0 (P0)

(Fig. 1b), similar to the expression pattern of EphrinB1 (Fig. 1c). At these stages, RGC axons have just arrived at the anterior end of the optic tectum and superior colliculus and have begun to be patterned on the pial surface. Sense controls are in Supplementary Fig. 1a and b. The *in situ* hybridization results were quantified by measuring the signal intensity using NIH Image J (Fig. 1d–f).

Both ephrinAs and ephrinB1 transcripts are found in the ventricular zone and their proteins are thought to be transported along radial glial cells to the pial surface of the superior colliculus^{10,12}. To test whether *Wnt3* protein is also transported to the pial surface, we performed western blotting using an anti-*Wnt3* antibody (Zymed Laboratories, Inc). Tectal tissues were isolated from the superficial layers of chick optic tectum from different medial-lateral positions and tissue lysates were equally loaded (1.5 µg protein per well) as indicated by the α -tubulin control (Fig. 1g). *Wnt3* protein was detected with a continuous medial-to-lateral decreasing gradient (M to L in Fig. 1g).

Biphasic and positional-dependent responses to Wnt3

To determine whether *Wnt3* can regulate the growth of RGC axons, we first examined their growth on polycarbonated filters coated with membrane fractions of HEK293 cells expressing *Wnt3*, taking advantage of the fact that *Wnt3* is highly hydrophobic and associates tightly with cell membranes^{13,14}. We found that *Wnt3*-transfected HEK293 cell membranes inhibited the growth of both dorsal and ventral mouse RGC axons at higher concentrations, and stimulated the growth of dorsal but not ventral RGC axons at lower concentrations (not shown). To obtain sufficient and consistent amounts of *Wnt3*, we overexpressed mouse *Wnt3* in SF9 cells using the Baculovirus system and tested the effects of different concentrations of affinity-purified *Wnt3* protein (coated with 0 ng ml^{−1}, 0.8 ng ml^{−1}, 4 ng ml^{−1} and 20 ng ml^{−1}) on chick RGC axons from six different dorsal-ventral positions (1–6 in Supplementary Fig. 2a). We found that at lower *Wnt3* concentrations, the growth of dorsal RGC axons

¹Department of Neurobiology, Pharmacology and Physiology, Committees on ²Developmental Biology and ³Neurobiology, University of Chicago, Chicago, Illinois 60637, USA.

*These authors contributed equally to this work.

was stimulated, whereas that of ventral axons was inhibited. At higher concentrations, both dorsal and ventral axon growth were inhibited (Fig. 1h). As a control, we performed the same series of experiments using control proteins from mock infected SF9 cells and found no attractive or repulsive effects (data not shown). Repulsion does not appear to be due to toxicity of the Wnt3 protein preparation because Wnt3 repulsion can be blocked by antibodies against Ryk (see below). Quantification is shown in Supplementary Fig. 2b. We tested retinal explants from different nasal and temporal positions and found that along the nasal–temporal axis, RGC axons did not display a graded responsiveness to Wnt3, suggesting that Wnt3 does not affect anterior–posterior topographic maps (not shown). Similar experiments were also performed with mouse retinal tissues from different dorsal ventral positions in different concentrations of Wnt3, which showed similar graded responsiveness (not shown).

Ryk is expressed in a ventral-to-dorsal decreasing gradient in RGCs

We found that Ryk, the mammalian homologue of a repulsive Wnt receptor, *Derailed*¹⁵, mediates repulsion of cortical axons by Wnts¹⁶. *In situ* hybridization showed that *Ryk* is expressed in a ventral-to-dorsal decreasing gradient in the RGCs of chick as early as E6

(Fig. 2a). In contrast, no dorsal–ventral gradient was observed with *frizzled5* at the same stage (Fig. 2b). At a later stage (E10 in chick), *Ryk* was also found expressed in a ventral-to-dorsal decreasing gradient in the RGC layer (Fig. 2c, d). No dorsal–ventral gradient of *frizzled5* expression was observed (Fig. 2e, f). Sense control is shown in Supplementary Fig. 1c and d. In mouse P0 retina, a similar dorsal–ventral gradient of *Ryk* mRNA was observed (Fig. 2g, h) and no dorsal–ventral gradient of *frizzled5* was detected (not shown). Because retinal tissues become much larger at E10 in chick and P0 in mouse, we showed higher magnification pictures from dorsal and ventral retina (Fig. 2c–h). Quantifications of *in situ* measurements are shown in Fig. 2i–k.

To determine the protein distribution of Ryk, we generated antibodies against the ectodomain of Ryk and performed immunohistochemistry. We first tested specificity of the Ryk antibodies by western blotting of E11.5 mouse embryonic extracts, and found that it recognized a highly specific band of the predicted size of 90 kD (Supplementary Fig. 4e)¹⁷. Immunostaining with chick E8 retina showed that Ryk protein is highly enriched on the axons of RGCs (Supplementary Fig. 3a–l). The immunoreactive signals on the RGC axon layer in ventral retina detected by Ryk antibodies (Supplementary Fig. 3a) co-localized with β -tubulin staining (E7 from Developmental

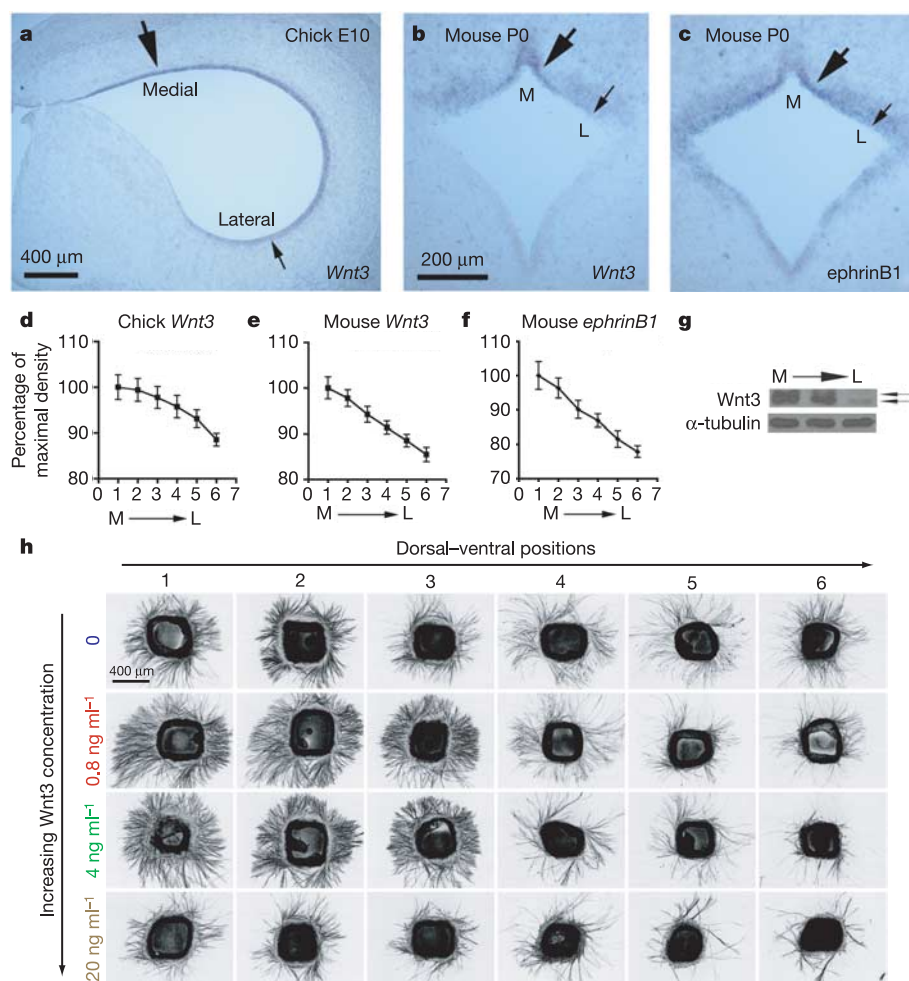


Figure 1 | Wnt3 expression is graded in vertebrate midbrain and Wnt3 differentially regulated retinal axon outgrowth along the dorsal-ventral axis. **a–f**, *In situ* hybridization demonstrating expression gradients in the ventricular zone (arrows) of *Wnt3* in E10 chick tectum (**a**) and P0 mouse superior colliculus (**b**), and ephrinB1 in P0 mouse superior colliculus (**c**), and quantification of *in situ* signal intensity (**d–f**). **g**, Western blot of membrane protein extracts from superficial optic tectum of E10 chick by anti-Wnt3 antibodies, showing medial-to-lateral decreasing gradient with

α -tubulin as a loading control. Multiple bands may represent post-translational modifications. **h**, Retina explant assay. E6 chick RGC explants from six dorsal–ventral positions (diagram in Supplementary Fig. 2a) were cultured on cover slips coated with control (0 ng ml⁻¹) and Wnt3 protein at three concentrations (0.8 ng ml⁻¹, 4 ng ml⁻¹ and 20 ng ml⁻¹), with quantification of outgrowth (Supplementary Fig. 2b). Scale bars: **a**, 400 μ m; **b**, **c**, 200 μ m; **h**, 400 μ m. M, medial; L, lateral. The error bars in all figures are s.e.m.

Biology Hybridoma Bank; Supplementary Fig. 3c), as indicated by the overlapping areas (yellow) in Supplementary Fig. 3d. Very little staining in the cell body was observed at this stage (indicated by DAPI nuclear staining in Supplementary Fig. 3b). Because axons from the entire dorsal–ventral axis project radially towards the optic disc to leave the eye, it is not possible to discern a potential gradient within the axon layer. At a later stage (E10), Ryk protein continues to be enriched in axons (Supplementary Fig. 3e–l), indicated by the overlapping areas with β -tubulin staining (yellow) in Supplementary Fig. 3h and i. At this stage, Ryk protein can be seen in the cell bodies of RGCs. Protein levels of Ryk on the ventral RGC cell bodies were found to be much higher than on the dorsal RGC cell bodies (Supplementary Fig. 3e, i), similar to the patterns of mRNA expression (Fig. 2c, d) (arrow in Supplementary Fig. 3e indicates axon layer).

We found that Ryk is a high-affinity receptor for Wnt3 with a K_d of 4.473 nM, and the K_d for the Wnt3–Frizzled5 interaction is 39.91 nM (Supplementary Fig. 4d). Frizzled3 has similar affinity to Wnt3 as Frizzled5 (not shown). Therefore, Ryk is a higher-affinity receptor for Wnt3 than Frizzled5 and Frizzled3.

We found that anti-Ryk antibodies ($50 \mu\text{g ml}^{-1}$) can specifically block Wnt3–Ryk binding (Supplementary Fig. 4g, $P < 0.0001$) but not Wnt3–Frizzled5 binding (Supplementary Fig. 4j, $P = 0.1069$). In contrast, sFRP2 ($0.2 \mu\text{g ml}^{-1}$) can block Wnt–Frizzled5 binding (Supplementary Fig. 4i, $P = 0.0045$) but cannot block Wnt–Ryk binding (Supplementary Fig. 4h, $P = 0.1094$). Similar results held true for Frizzled3 (not shown). The mechanism of Wnt–Ryk binding may be different from that of Wnt–Frizzled. The Wnt-binding domain in the Frizzled protein is the cysteine rich domain (CRD)¹⁸ and the domain in Ryk for Wnt binding is the structurally unrelated Wnt-inhibitory factor (WIF) domain^{19,20}. Because of the differential blocking effect, sFRP2 protein and anti-Ryk antibodies can be used to tease apart the function of Frizzled and Ryk, by specifically blocking the binding of Wnt3 to Frizzled(s) or Ryk, respectively.

Ryk mediates inhibition and Frizzleds mediate stimulation

We tested whether anti-Ryk antibodies can block the Wnt3 effects on dorsal and ventral axons at low and high concentrations (0.8 ng ml^{-1} and 20 ng ml^{-1}) (Fig. 3). For dorsal retinal explants, we dissected

and cultured the retinal tissue from position 2, as indicated in Supplementary Fig. 2a. For ventral explants, we used position 5, as shown in Supplementary Fig. 2a. We found that the inhibition of ventral axons by Wnt3 at both concentrations, and that of dorsal axons at 20 ng ml^{-1} , can be blocked by the Ryk antibodies ($50 \mu\text{g ml}^{-1}$). However, Wnt3-mediated stimulation of dorsal axons at 0.8 ng ml^{-1} cannot be blocked by the anti-Ryk antibodies (Fig. 3a).

Several lines of evidence suggest that the Ryk antibodies are highly specific. The specificity of the Ryk antibodies was first tested by western blot (Supplementary Fig. 4e). In binding assays, the Ryk antibodies only blocked the binding of Wnt3 to Ryk (Supplementary Fig. 4g) but not the binding to Frizzled5, suggesting that the Ryk antibodies do not cross-react with Frizzled5 (Supplementary Fig. 4j) or Frizzled3 (not shown). In addition, the Ryk antibodies did not block the stimulation of dorsal RGC axons by low concentrations of Wnt3 (Fig. 3a). This result itself also serves as an internal control for the specificity of the antibodies. To further test the specificity of the anti-Ryk antibodies, we performed two additional control experiments, whereby we found that Ryk antibodies do not block Slit2 repulsion and Wnt4 attraction of post-crossing commissural axons (published in Fig. 4c–f in ref. 16).

Neither preimmune nor anti-Ryk postimmune sera had any effect on the growth of RGC axons in the absence of Wnt3 (blue bars in Fig. 3b, c), suggesting that the rabbit sera did not promote growth in general. It is interesting that the Ryk antibodies allowed for Wnt3-mediated growth promotion (red bars in pre-immune and anti-Ryk in Fig. 3b and c). It is possible that the growth cones of dorsal RGC axons have both attractive and repulsive signalling pathways. When repulsion is inhibited by the Ryk antibodies, the attractive pathway takes over and shifts the balance. It should be noted that although anti-Ryk antibodies completely blocked the inhibition of dorsal RGC axons at a higher Wnt3 concentration (20 ng ml^{-1}) (green bars in preimmune and anti-Ryk in Fig. 3b), they did not completely block Wnt3 inhibition of ventral axons, particularly at a high concentration (up to 80% for 0.8 ng ml^{-1} and up to 55% for 20 ng ml^{-1}) (red and green bars in preimmune and anti-Ryk in Fig. 3c). This could be because the ventral axons express more Ryk or the antibodies have limited efficacy. One also cannot exclude the possibility that the

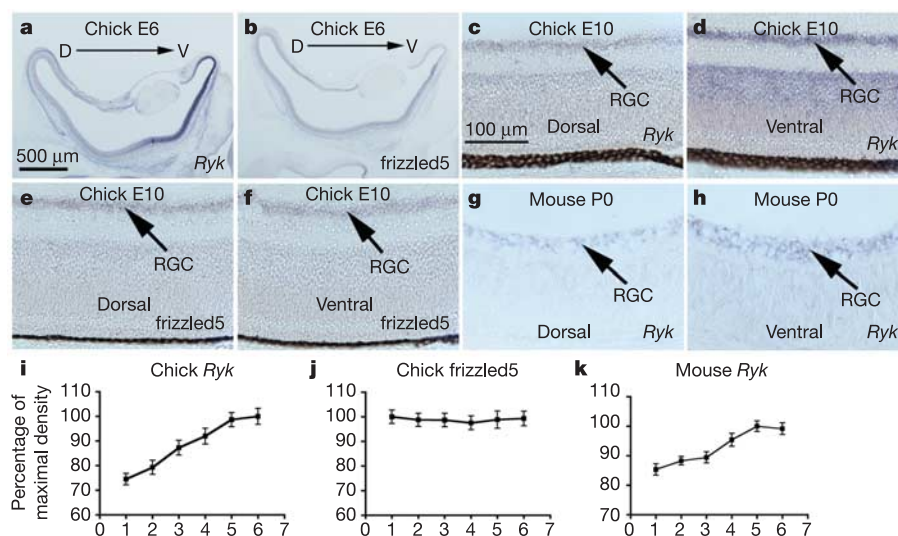


Figure 2 | Graded expression of Ryk in chick and mouse retinal ganglion cells. a–k, *In situ* hybridization and signal intensity quantification. Ryk mRNA is expressed in a ventral-to-dorsal decreasing gradient in E6 chick retina (a), while *frizzled5* transcript is expressed evenly (b). Higher magnification of *Ryk* *in situ* hybridization in E10 dorsal (c) and ventral (d) chick retina and *in situ* hybridization of *frizzled5* in dorsal (e) and ventral (f)

E10 chick retina. g, h, Higher magnification of *Ryk* *in situ* hybridization in dorsal (g) and ventral (h) P0 mouse retina. Quantification of *in situ* hybridization signal intensity along the dorsal–ventral axis of *Ryk* (i) and *frizzled5* (j) in E10 chick retina, and *Ryk* expression in P0 mouse retina (k). Scale bars: a, b, 500 μm ; c–h, 100 μm . Numbers 1–6 represent six different positions along the dorsal–ventral axis of the retina (Supplementary Fig. 2a).

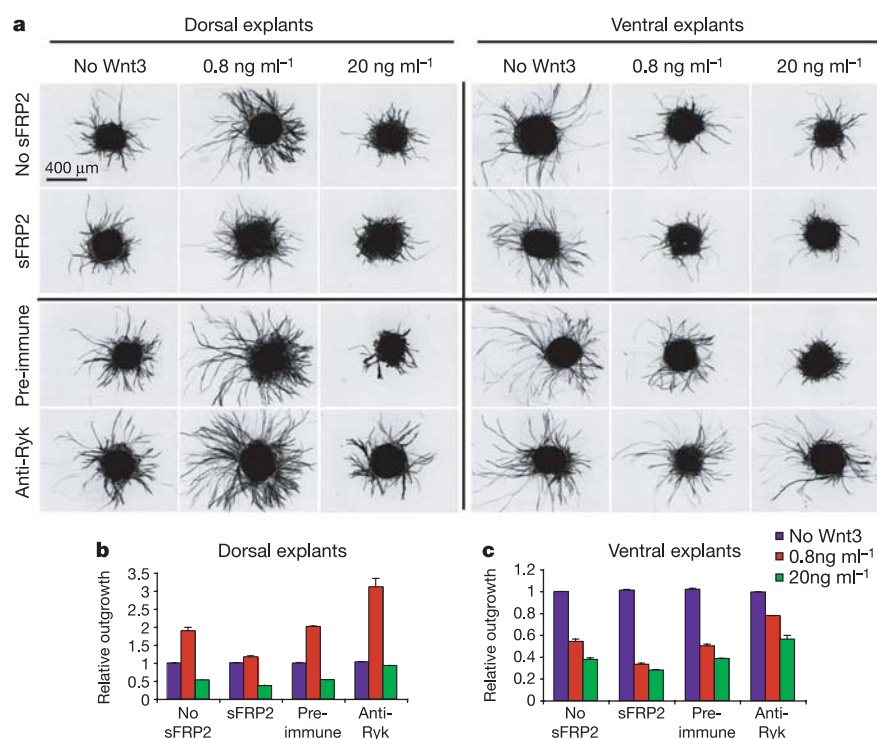


Figure 3 | Wnt3 inhibits retinal ganglion cell axons via Ryk and stimulates retinal ganglion cell axons via Frizzled(s). **a**, Chick E6 retinal explants from both dorsal and ventral retina were cultured on cover slips coated with low or high concentrations of Wnt3 recombinant protein in the presence of preimmune or anti-Ryk antibodies or in the presence of sFRP2. Wnt3 inhibition of ventral explants at both concentrations and of dorsal explants

at high concentrations can be blocked by anti-Ryk antibodies but cannot be blocked by sFRP2. Stimulation to dorsal retinal explants at lower Wnt3 concentrations can be blocked by sFRP2 but cannot be blocked by anti-Ryk antibodies. **b**, **c**, Quantification of dorsal (**b**) and ventral (**c**) explants in the Ryk antibody blocking and sFRP2 blocking experiments at different Wnt3 concentrations. Scale bar in **a**, 400 μm.

ventral axons contain an unknown repulsive Wnt3 guidance receptor other than Ryk.

To address whether Frizzled(s) mediate stimulation or inhibition by Wnt3, we tested the purified sFRP2 protein at two concentrations of Wnt3. The stimulation of dorsal explants at low concentrations of Wnt3 (0.8 ng ml⁻¹) can be blocked by sFRP2 (0.2 μg ml⁻¹),

whereas the inhibition of ventral and dorsal axons at a higher Wnt3 concentration (20 ng ml⁻¹) cannot be blocked by sFRP2 (Fig. 3a).

Ectopic Wnt3 expression in tectum repelled the termination zones

To test whether Wnt3 repels ventral axon termination zones *in vivo*,

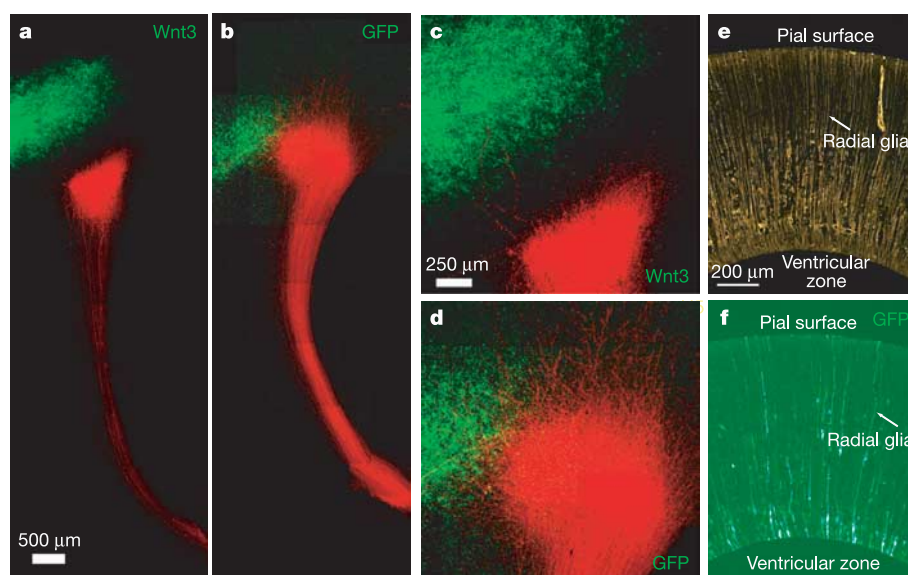


Figure 4 | Termination zone of abnormalities in tectum ectopically expressing Wnt3. **a**, RGC axon termination zones are repelled by Wnt3 expression (green area) ($n = 7$). **b**, Normal termination zone formation of RGC axons in the tectal area expressing GFP control ($n = 6$). **c**, Higher magnification of **a**. **d**, Higher magnification of **b**. **e**, Immunostaining of chick

E14 optic tectum with an antibody against a radial glial marker, H5, showing the presence of radial glial fibres. **f**, Immunostaining of chick E14 optic tectum electroporated with a GFP construct in the ventricular zone, showing that GFP protein was transported to the pial surface along the radial glial cells. Scale bars: **a**, **b**, 500 μm; **c**, **d**, 250 μm; **e**, **f**, 200 μm.

we overexpressed Wnt3 by electroporation in the ventricular zone of chick optic tectum. Because the tectum is patterned by E3, expressing Wnt3 four days later (E7) should not alter the patterning of tectum²¹. In addition, we tested the expression pattern of ephrinB1 in the chick tectum electroporated with Wnt3 and found that the normal medial–lateral gradient of ephrinB1 was not altered (Supplementary Fig. 5b).

We electroporated a Wnt3 expression construct at E7, traced RGC axon termini with DiI injection at E13, and harvested tecta on E14 (Supplementary Fig. 5a). We found that the RGC axon termination zone, labelled with DiI, was repelled by ectopic Wnt3 (Fig. 4a, c, $n = 7$) as compared to the green fluorescent protein (GFP) only (Fig. 4b, d, $n = 6$), confirming a repulsive role of Wnt3 in guiding the termination of RGC axons. Immunostaining with the radial glial marker, H5, confirmed that radial glial fibres are present in chick tectum at this stage (Fig. 4e). We also stained a slice of chick optic tectum electroporated with a GFP construct in the ventricular zone with a GFP antibody, and found that many radial glial cells were expressing GFP. GFP was also detected all the way on the pial surface, confirming that proteins expressed in the radial glial cells in the ventricular zone introduced by electroporation can be transported to the pial surface (Fig. 4f).

Dominant-negative Ryk caused a medial shift of the termination zone

Because Wnt3 is expressed in a medial-to-lateral decreasing gradient, we hypothesize that Wnt3 directs axons laterally to counteract EphrinB1, which promotes medially directed growth. Therefore, we predict that by blocking Wnt3–Ryk function, the termination zone should shift medially. Wnt3 knockout mice fail in early embryonic patterning and Wnt3 is also important for early nervous system development, making it impossible to examine the function of Wnt3 *in vivo* using the conventional knockout mouse approach²². Ryk knockout mice die at birth²³ and cannot be used to evaluate the function of Wnt3–Ryk in mice as termination zones form at postnatal day 8¹⁰. To circumvent this difficulty, we generated a dominant-negative form of Ryk and expressed it in chick RGCs in the dorsal aspect of retina by *in ovo* electroporation at E5 (Supplementary Fig. 6a, right panel). The truncated Ryk protein only contains Ryk ectodomain and the transmembrane domain, missing the intracellular domain. The intracellular domain was shown to be required for axon guidance in the fly homologue, *Derailed*¹⁵. It is unlikely that the expression of the truncated Ryk in the retina at E5 will affect dorsal–ventral fate mapping. First, the intracellular domain of Ryk is dispensable in cell fate determination, as a truncated Ryk, including only the extracellular and the transmembrane domains, can rescue cell fate phenotypes in *Caenorhabditis elegans*²⁴. Second, chick retinal cell fate is determined at an earlier stage (Hamilton–Hamburger stage 10 or E1.5), and after stage 10, fate is already determined, as suggested by reported studies²¹. For example, cells from the anterior eye anlage were manipulated to form the posterior anlage, and they maintained the anterior identity and selected a tectal target consistent with the anterior origin²¹. Third, we examined the expression patterns of cell differentiation markers, such as EphrinB1 and EphB2, and found that their normal graded expression patterns were not affected (Supplementary Fig. 6b).

To visualize RGC axons, we co-electroporated a mixture of the dominant-negative Ryk and a cytomegalovirus (CMV)–GFP construct at a 3:1 ratio (Ryk DN:GFP). Mixing these two constructs allows us to determine the width of the termination zone (because some of the RGC axons will express the GFP control only) and the relative medial–lateral position of the termination zone when the Ryk dominant-negative construct is expressed. We elected to express the Ryk dominant-negative construct in the dorsal retina because Ryk is expressed at lower levels there and, therefore, it may be easier to block the endogenous Ryk function by the truncated Ryk. In addition, the

dorsal axons normally target the lateral optic tectum, so we can test the hypothesis that a Wnt–Ryk interaction mediates lateral-directed axon termination.

We found that RGC axons, co-electroporated with the dominant-negative Ryk construct, formed wide termination zones that extended more medially (Fig. 5b) compared to GFP control (Fig. 5a). These termination zones typically expand to at least twice the normal size, and the medial extreme of the termination zone extended widely towards the dorsal midline and only shifted medially.

Dominant-negative Ryk eliminated lateral-directed interstitial branches

To further characterize the topographic map shift and analyse interstitial branches, we created a construct with CMV-enhanced chick β -actin promoter driving dominant-negative Ryk followed by

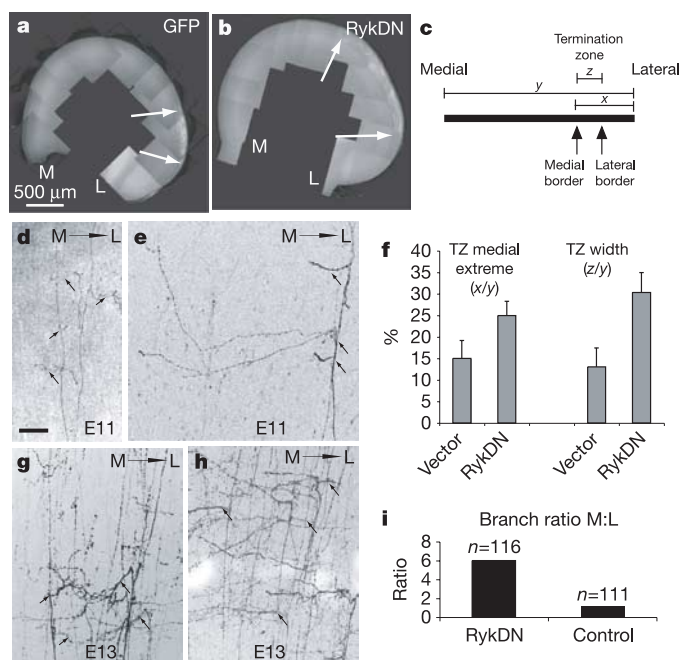


Figure 5 | Ryk is required for normal medial–lateral patterning of RGC axon termination. **a**, Normal termination zone of dorsal retinal ganglion cell axons in chick tectum visualized by eGFP in a control construct (CMV–eGFP) in 300 μ m vibratome sections perpendicular to the anterior–posterior axis of the tecta contralateral to retinas. eGFP was directly visualized by fluorescence. **b**, Much diffused termination zone (TZ) of dorsal RGC axons expressing a dominant-negative (DN) Ryk construct. The termination zone shifted medially. **c**, Diagram showing the quantification method. Dorsal RGC axons terminate at the lateral edge of the tectum. Termination zone (z) is the area with eGFP signal observed in the vibratome section. The TZ width is defined as the ratio of the length of the termination zone (z) over the entire length of the tectum along the medial–lateral axis (y). The TZ medial extreme is defined as the ratio of the distance from the lateral edge to the medial border of the termination zone (x) over the distance of the entire medial–lateral axis (y). The termination zone shifted medially and expanded in size when dominant-negative Ryk was expressed in the dorsal retinal ganglion cells. **d, g**, Both medial- and lateral-directed interstitial branches were observed emerging from the primary RGC axons expressing a GFP control (n of tecta = 20). Leftward arrows indicate medial-directed branches. Rightward arrows indicate lateral-directed branches. **e, h**, Interstitial branches of retinal ganglion cells in the tectum pointed medially from primary RGC axons expressing dominant-negative Ryk, visualized by an IRES–eGFP marker (n of tecta = 27). Virtually no lateral-directed interstitial branches were observed at the termination zone. **f**, Quantification of medial spread and width of termination zone. **g**, Quantification of ratio of medial- and lateral-directed interstitial branches in RGC axons expressing dominant-negative Ryk and a eGFP control.

internal ribosomal entry site (IRES)-eGFP. Using this construct, all green axons should express the dominant-negative Ryk, allowing us to observe details of branch formation from the primary RGC axon shafts. This construct was introduced into RGC cells at E7 by electroporation, and the axon projections in the whole mount tectum were analysed from E11 to E13 (Fig. 5d–h). We found that in Ryk dominant negative-expressing axons, very few lateral-directed branches were observed at the termination zone and only medial-directed branches were present (Fig. 5e, h) and they were typically longer than in GFP control-electroporated axons, which displayed almost equal length and frequency of interstitial branches of both directions (Fig. 5d, g). These long medial branches eventually formed multiple smaller termination zones medial to their appropriate positions along the medial–lateral axis of the tectum (not shown).

Discussion

Our study proposes that Wnt3 acts as a lateral mapping force in the optic tectum to counterbalance the EphrinB1–EphB interaction, which is a medial-directed mapping force¹⁰ (Fig. 6a). The Wnt3 and EphrinB1 signalling pathways are probably independent of each other, as blocking of Wnt–Ryk function allows axons to still respond to the EphrinB1–EphB function, causing termination zones to shift medially (Fig. 6b), and the termination zones to shift laterally in EphB2/B3 double knockout mice (Fig. 6c)¹⁰. However, one cannot exclude the possibility that these two signalling pathways may modulate each other's activity in the same growth cone. Wnt3 is the only Wnt gene that displays a continuous medial–lateral graded expression in the ventricular zone of the tectum and superior colliculus. Initial mapping of the retinal axons occurs at the pial surface, which contains molecular cues such as the EphrinAs and EphrinB1 that are synthesized by radial glial cells and transported along the radial glia to the pial surface. However, our current study cannot exclude the possible roles of other Wnts, which might contribute to map formation through diffusion from other layers of the tectum or superior colliculus. Owing to technical limitations of *in ovo* electroporation, it is not feasible to systematically examine

map formation throughout the entire medial–lateral axis to confirm that these two mapping forces function uniformly, or to determine whether there are additional mapping forces. More sophisticated genetic manipulations will be needed to address these questions.

The graded and biphasic activity of Wnt3 in different dorsal–ventral retinal axons is reminiscent of EphrinAs along the anterior–posterior axis²⁵. The differential responsiveness to EphrinAs was proposed to be caused by repulsion versus adhesion, as repulsion requires cleaved EphrinAs and adhesion requires glycosylphosphatidylinositol (GPI)-linked ephrinAs²⁵. Our study suggests an alternative model such that a repulsive Wnt–Ryk pathway competes with an attractive Wnt–Frizzled interaction to titrate the response to Wnt3 protein at different concentrations. The graded response may be determined by the gradient of Ryk expression in the dorsal–ventral axis. Ventral RGCs express more Ryk, whereas dorsal RGCs have less Ryk. Expression of frizzled5 appears to be even along the dorsal–ventral axis. The net outcome of this competition is varied in a graded fashion in growth cones along the dorsal–ventral axis, and thus determines the topographic connections.

Neurons typically connect to multiple targets and even different brain centres. Multiple forms of axon branches are observed, including collateral and interstitial branches. During retinotectal topographic mapping, RGCs project interstitial branches. These interstitial branches extend medially or laterally towards their future termination zone²⁶. Once interstitial branches reach their termination zone, they form elaborate termini and the axons that have overshoot are pruned back (in chick and mice). Therefore, the initial direction and growth of interstitial branches influence the position of the termination zone. We found that the Wnt–Ryk pathway is required for the laterally directed interstitial branches *in vivo*. Blocking Ryk by a truncated Ryk receptor eliminated nearly all laterally directed branches, leaving only the medially directed branches, which are unusually long. In contrast, in EphB2 and EphB3 double knockout mice, interstitial branches were found preferentially directed laterally¹⁰. Therefore, Wnt3 and EphrinB1 are opposing guidance activities for regulating interstitial branches in medial–lateral retinotectal mapping (Fig. 6).

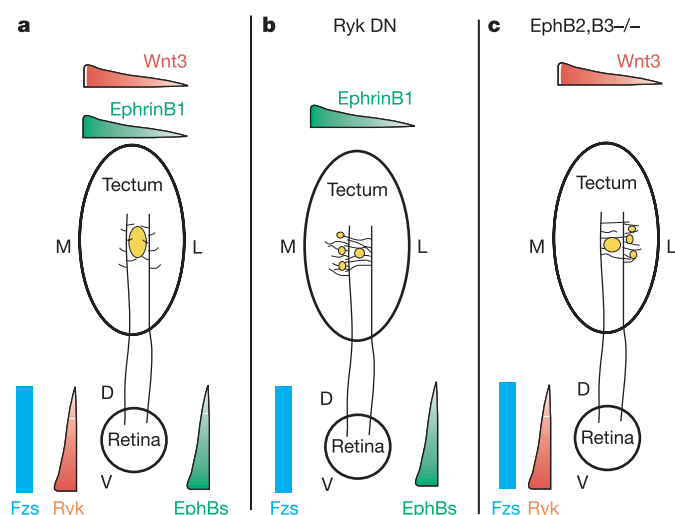


Figure 6 | Model of two counterbalancing forces for medial–lateral map formation. **a**, Wnt3 is a lateral mapping force, which is mediated by Ryk and Frizzled. Ventral axons are repelled by Wnt3, but dorsal axons are attracted by low Wnt3 and repelled by high Wnt3. Although the effect is biphasic along the dorsal–ventral axis of retina, the overall function of Wnt signalling is to drive interstitial branches laterally. The medial mapping force is EphrinB1 read by EphB2 and B3 via an attractive mechanism. **b**, When Wnt3 repulsion is blocked by a Ryk dominant-negative construct, EphrinB1–EphBs signalling causes interstitial branches to project medially only, causing a medial shift of the map. **c**, When EphrinB1–EphBs signalling was eliminated, interstitial branches project laterally, as shown in ref. 10.

METHODS

Retinal explant cultures. Retinal explant cultures were performed according to a modified procedure from a previously described method²⁵. For details, see Supplementary Methods.

In ovo electroporation. For Wnt3 ectopic overexpression, E7 chick tecta were electroporated with a CMV-Wnt3 expression construct mixed with a CMV-enhanced GFP (eGFP) construct (Wnt3:eGFP = 3:1) to visualize the electroporated area, and at E13 RGC axons were labelled with a focal injection of Di-I (diagram in Supplementary Fig. 5a). Tecta were then harvested at E14 for analyses of retinotectal projections (Fig. 4).

To characterize the termination zone of RGC axons, E5 chick retinas were electroporated in dorsal retina with a Ryk dominant-negative construct mixed with eGFP construct (3:1 ratio) (Supplementary Fig. 6a). On E14, contralateral tecta were harvested, and sectioned (250–300 μ m thick) perpendicular to the anterior–posterior axis with a vibratome. Sections were photographed on an epifluorescence microscope (Fig. 5a, b). Results were quantified from four Ryk dominant-negative and three control experiments (Fig. 5c, f). The relative medial–lateral positions of termination zones are dependent on the dorsal–ventral position of electroporation in the retina. Therefore, the results of the medial extreme (the most medial border) of the termination zone have higher system error. However, the results on the width of termination zones are independent of the dorsal–ventral positions of electroporation and therefore are less prone to variations caused by the site of electroporation.

For analyses of interstitial branches, E7 chick retinas were electroporated with pCIG Ryk dominant-negative IRES GFP and contralateral tecta were harvested from E11 to E13 (Supplementary Fig. 6a). Tecta were flat mounted on glass slides and photographed with a confocal microscope (Fig. 5d, e, g, h). Results were quantified by counting 116 branches in 27 tecta for the Ryk dominant-negative construct and 111 branches in 20 tecta for the GFP only control (Fig. 5i).

Cloning and constructs. Detailed cloning and construct information can be found in Supplementary Methods.

In situ hybridization, immunohistochemistry and western blot. *In situ* hybridization and immunohistochemistry were performed as previously described¹¹. For quantification of *in situ* signal density, digital images of *in situ* hybridizations were taken and resized to 600 × 450 pixels. The positive greyscale signals along the dorsal–ventral axis of the RGC layer of retina or the medial–lateral extent in the ventricular zone of the chick optic tectum and mouse superior colliculus were quantified with NIH Image by plot profile. The average signal density of each segment, retinal RGC layer or the ventricular zone of tectum or superior colliculus was divided into six equal segments and calculated. Data from five sections were collected and averaged. Graphs were created using GraphPad Prism after data were normalized by defining the largest value as 100%. Owing to the nonlinear nature of the enzyme reaction, this quantification represents the direction of the *in vivo* gradient but may not reflect the actual steepness.

Polyclonal anti-Wnt3 antibodies were purchased from Zymed Laboratories, Inc and used at 1 µg ml⁻¹ for western blot (1:100 dilution). H5 (1:100 dilution for immunohistochemistry) and E7 (1:500 dilution for immunohistochemistry) were monoclonal antibodies purchased from Developmental Biology Hybridoma Bank. Purified anti-Ryk antibodies were diluted 1:500 for immunostaining (Supplementary Fig. 3) and 1:1,000 for western blot (Supplementary Fig. 4e).

Wnt receptor binding assays. The protocol for binding assay was performed as previously published^{27,28}. Details of methods can be found in Supplementary Methods.

Received 7 July; accepted 19 October 2005.

Published online 9 November 2005.

- Prestige, M. C. & Willshaw, D. J. On a role for competition in the formation of patterned neural connexions. *Proc. R. Soc. Lond. B* **190**, 77–98 (1975).
- Gierer, A. Model for the retino-tectal projection. *Proc. R. Soc. Lond. B* **218**, 77–93 (1983).
- Fraser, S. E. & Hunt, R. K. Retinotectal specificity: models and experiments in search of a mapping function. *Annu. Rev. Neurosci.* **3**, 319–352 (1980).
- Fraser, S. E. & Perkel, D. H. Competitive and positional cues in the patterning of nerve connections. *J. Neurobiol.* **21**, 51–72 (1990).
- Flanagan, J. G. & Vanderhaeghen, P. The ephrins and Eph receptors in neural development. *Annu. Rev. Neurosci.* **21**, 309–345 (1998).
- Frisen, J. et al. Ephrin-A5 (AL-1/RAGS) is essential for proper retinal axon guidance and topographic mapping in the mammalian visual system. *Neuron* **20**, 235–243 (1998).
- Feldheim, D. A. et al. Topographic guidance labels in a sensory projection to the forebrain. *Neuron* **21**, 1303–1313 (1998).
- Feldheim, D. A. et al. Genetic analysis of ephrin-A2 and ephrin-A5 shows their requirement in multiple aspects of retinocollicular mapping. *Neuron* **25**, 563–574 (2000).
- Mann, F., Ray, S., Harris, W. & Holt, C. Topographic mapping in dorsoventral axis of the *Xenopus* retinotectal system depends on signalling through ephrin-B ligands. *Neuron* **35**, 461–473 (2002).
- Hindges, R., McLaughlin, T., Genoud, N., Henkemeyer, M. & O'Leary, D. D. EphB forward signalling controls directional branch extension and arborization required for dorsal–ventral retinotopic mapping. *Neuron* **35**, 475–487 (2002).
- Lyuksyutova, A. I. et al. Anterior–posterior guidance of commissural axons by Wnt-frizzled signalling. *Science* **302**, 1984–1988 (2003).
- Drescher, U. et al. *In vitro* guidance of retinal ganglion cell axons by RAGS, a 25 kDa tectal protein related to ligands for Eph receptor tyrosine kinases. *Cell* **82**, 359–370 (1995).
- Smolich, B. D., McMahon, J. A., McMahon, A. P. & Papkoff, J. Wnt family proteins are secreted and associated with the cell surface. *Mol. Biol. Cell* **4**, 1267–1275 (1993).
- Willert, K. et al. Wnt proteins are lipid-modified and can act as stem cell growth factors. *Nature* **423**, 448–452 (2003).
- Yoshikawa, S., McKinnon, R. D., Kokel, M. & Thomas, J. B. Wnt-mediated axon guidance via the *Drosophila* Derailed receptor. *Nature* **422**, 583–588 (2003).
- Liu, Y. et al. Ryk-mediated Wnt repulsion regulates posterior-directed growth of corticospinal tract. *Nature Neurosci.* **8**, 1151–1159 (2005).
- Hovens, C. M. et al. RYK, a receptor tyrosine kinase-related molecule with unusual kinase domain motifs. *Proc. Natl Acad. Sci. USA* **89**, 11818–11822 (1992).
- Dann, C. E. et al. Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. *Nature* **412**, 86–90 (2001).
- Hsieh, J. C., Rattner, A., Smallwood, P. M. & Nathans, J. Biochemical characterization of Wnt-frizzled interactions using a soluble, biologically active vertebrate Wnt protein. *Proc. Natl Acad. Sci. USA* **96**, 3546–3551 (1999).
- Patthy, L. The WIF module. *Trends Biochem. Sci.* **25**, 12–13 (2000).
- Mey, J. & Thanos, S. Development of the visual system of the chick. I. Cell differentiation and histogenesis. *Brain Res. Rev.* **32**, 343–379 (2000).
- Liu, P. et al. Requirement for Wnt3 in vertebrate axis formation. *Nature Genet.* **22**, 361–365 (1999).
- Halford, M. M. et al. Ryk-deficient mice exhibit craniofacial defects associated with perturbed Eph receptor crosstalk. *Nature Genet.* **25**, 414–418 (2000).
- Inoue, T. et al. *C. elegans* LIN-18 is a Ryk ortholog and functions in parallel to LIN-17/Frizzled in Wnt signalling. *Cell* **118**, 795–806 (2004).
- Hansen, M. J., Dallal, G. E. & Flanagan, J. G. Retinal axon response to ephrin-as shows a graded, concentration-dependent transition from growth promotion to inhibition. *Neuron* **42**, 717–730 (2004).
- McLaughlin, T., Hindges, R. & O'Leary, D. D. Regulation of axial patterning of the retina and its topographic mapping in the brain. *Curr. Opin. Neurobiol.* **13**, 57–69 (2003).
- Flanagan, J. G. & Leder, P. The kit ligand: a cell surface molecule altered in steel mutant fibroblasts. *Cell* **63**, 185–194 (1990).
- Cheng, H. J. & Flanagan, J. G. Identification and cloning of ELF-1, a developmentally expressed ligand for the Mek4 and Sek receptor tyrosine kinases. *Cell* **79**, 157–168 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Alfred Sloan Foundation, the Schweppe Foundation and NINDS. We thank F. Polleux for the pCIG2 vector (CMV-enhanced β-actin promoter with IRES GFP marker) and A. G. Fenstermaker for critical reading of the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.Z. (yzou@bsd.uchicago.edu).

ARTICLES

Structure of the parainfluenza virus 5 F protein in its metastable, prefusion conformation

Hsien-Sheng Yin^{1,2}, Xiaolin Wen², Reay G. Paterson², Robert A. Lamb^{1,2} & Theodore S. Jardetzky²

Enveloped viruses have evolved complex glycoprotein machinery that drives the fusion of viral and cellular membranes, permitting entry of the viral genome into the cell. For the paramyxoviruses, the fusion (F) protein catalyses this membrane merger and entry step, and it has been postulated that the F protein undergoes complex refolding during this process. Here we report the crystal structure of the parainfluenza virus 5 F protein in its prefusion conformation, stabilized by the addition of a carboxy-terminal trimerization domain. The structure of the F protein shows that there are profound conformational differences between the pre- and postfusion states, involving transformations in secondary and tertiary structure. The positions and structural transitions of key parts of the fusion machinery, including the hydrophobic fusion peptide and two helical heptad repeat regions, clarify the mechanism of membrane fusion mediated by the F protein.

The Paramyxoviridae are enveloped viruses that include, among others, mumps virus, measles virus, Sendai virus, Newcastle disease virus (NDV), human respiratory syncytial virus (RSV), parainfluenza virus 5 (SV5) and human parainfluenza viruses 1–4 (hPIV)¹. Many members of this viral family are significant human and animal pathogens, and newly emergent deadly paramyxoviruses (Nipah and Hendra viruses^{2,3}) have been identified.

The paramyxoviruses, like other enveloped viruses such as influenza virus and HIV, require fusion of the viral and cellular membranes to enter the host cell. Two viral glycoproteins are key to this process: a variable attachment protein (HN, H or G) and a more conserved fusion (F) protein¹. The attachment proteins interact with different cellular receptors. For example, SV5 HN protein binds sialic acid, measles virus H protein interacts with CD46 or CDw150/SLAM (refs 4, 5), Nipah and Hendra virus G proteins bind to Ephrin B2 (refs 6, 7) and RSV G protein binds heparin sulphate⁸. Although the cellular receptors differ, in most paramyxoviruses the homotypic attachment protein is required to trigger F-mediated membrane fusion at the right time and right place^{9,10}. F is not activated by the low pH found in the endosome¹¹.

F is thought to drive membrane fusion by coupling irreversible protein refolding to membrane juxtaposition, by initially folding into a metastable form that subsequently undergoes discrete conformational changes to a lower energy state^{9,10}. F assembles into homotrimers that are proteolytically cleaved, priming the protein for membrane fusion (Fig. 1), similar to the influenza virus haemagglutinin¹² (HA) and other class I viral fusion proteins such as HIV Env, Ebola virus GP and SARS coronavirus S (refs 12, 13). The uncleaved precursor (F₀) is processed into a larger carboxy-terminal fragment (F₁) and a smaller amino-terminal fragment (F₂). F₁ contains a hydrophobic fusion peptide at its N terminus and two hydrophobic, heptad repeat regions (HRA and HRB). HRA is immediately adjacent to the fusion peptide and HRB is proximal to the transmembrane domain, with about 250 residues separating the two.

After activation, F inserts its fusion peptide into target membranes¹⁴, forming transient intermediates^{15,16} that can be inhibited by HRA- and HRB-derived peptides. Subsequent refolding and assembly of HRA and HRB into a six-helix bundle (6HB) occurs, placing the fusion peptides and the transmembrane domains in proximity^{13,17}. Formation of the 6HB and the associated free-energy change are tightly linked to merger of the viral and cellular membranes^{16,18}. The isolated F 6HB structure, generated from HRA and HRB peptides^{17,19}, is stable up to 100 °C and is thought to represent the lowest-energy conformation of the protein after membrane fusion. We previously reported the structure of the uncleaved, secreted hPIV3 F₀ ectodomain (solF₀), truncated before the transmembrane domain²⁰. Unexpectedly, this structure contains the postfusion 6HB, indicating that F protein cleavage is not required to attain the postfusion conformation and that the F transmembrane domain and/or the cytoplasmic tail are important for the folding to, or stability of, the prefusion metastable state²⁰. It remained unclear to what extent the F pre- and postfusion conformations differ^{21,22} and how these are linked to membrane fusion.

We now report the structure of the SV5 F protein in the prefusion conformation. The structure contains a globular head attached to a trimeric coiled-coil stalk formed by the C-terminal HRB region. The globular head contains three domains (DI–DIII) identified previously^{20,21}. DI and DII reposition as rigid modules during the conformational transition, and the main refolding occurs in DIII. Core structural elements of DIII act as a scaffold for the folding of HRA, preventing its assembly into the postfusion helical conformation. The fusion peptides at the N termini of HRA segments are sequestered between adjacent subunits, and cleavage/activation sites are exposed at the protein surface. None of the intersubunit contacts are conserved in the pre- and postfusion forms. The SV5 F structure provides a model for the stepwise induction of membrane fusion by paramyxoviruses and shows how several sequence elements have distinct structural roles in the pre- and postfusion conformations.

¹Howard Hughes Medical Institute, and ²Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, Illinois 60208-3500, USA.

Crystallization and structure determination

Previous attempts to determine the structure of the prefusion conformation of F were confounded by spontaneous folding of the anchorless, secreted hPIV3 and NDV F proteins to the trimeric, postfusion state^{20–22}. Similarly truncated SV5 F does not trimerize efficiently and therefore we appended an engineered, trimeric coiled-coil domain²³ (GCNt) to HRB to mimic the transmembrane domains (Fig. 1). The resulting F protein (F-GCNt) assembles into well-defined trimers (Supplementary Fig. 6). Appending GCNt to the F cytoplasmic tail stabilizes the trimer and reduces its fusogenicity²⁴. Conceptually related constructs have been reported for the HIV Env and influenza virus HA proteins^{25–28}.

F-GCNt crystals were obtained that diffract X-rays to 2.85 Å (Table 1 and Supplementary Tables 2 and 3). The structure could not be solved with the NDV or hPIV3 F structures and was determined by isomorphous replacement methods, combined with three-fold non-crystallographic symmetry (NCS) averaging (see

Methods and Supplementary Information). The SV5 F-GCNt model was refined to a final R_{free} of 26.1%.

Structure of the F-GCNt trimer

The F-GCNt trimer has a large globular head attached to a three-helix coiled-coil stalk formed by HRB (Fig. 1b–e). The GCNt trimer is located at the C-terminal end of the stalk, orienting the head away from the viral membrane. The top of the head region has three prominent spikes formed by two pairs of loops (60–65 and 178–185) that project upwards from the globular domains in each subunit.

The F-GCNt head contains three domains (DI–DIII) per subunit that extend around the trimer axis, making extensive intersubunit contacts. A large cavity is present at the base of the head, with the bottom and sides formed by DI and DII. DIII (residues 42–278) covers the top of the cavity and includes the prominent spikes, HRA and the fusion peptide (Fig. 1 and Supplementary Fig. 7). At the C terminus of DII, an extended linker to HRB wraps around the outside of the trimer and into the centre of the base of the head where the stalk begins. Electron density for the HRB linker was weaker, suggesting flexibility of this region. The structure has three lateral vertices projecting from the trimer axis, exposing the cleavage/activation sites adjacent to the fusion peptides (Fig. 1c, d). Helices line the central three-fold axis at the top and bottom of the trimer. In DIII, two sets of six helices form rings covering the top of the head, and the HRB three-helix bundle seals the bottom (Fig. 1d).

Structure and location of the fusion peptide

In the postfusion structure of hPIV3 solF0, the fusion peptide could not be modelled and was most probably located on the exterior of the F trimer stalk²⁰. In the prefusion SV5 F structure, by contrast, strong electron density is observed for the hydrophobic fusion peptide (residues 103–128), which is wedged between two subunits of the trimer (Fig. 1e). The N-terminal end of the fusion peptide is exposed at the F surface and then proceeds inwards, becoming more buried from solvent.

The fusion peptide adopts a partly extended, partly β -sheet and partly α -helical conformation and is sandwiched between DIII of its own subunit and DII of another. Residues 107–117 pack against the hydrophobic edge of the neighbouring DII, interacting with the first (A) and last (G) strands of the immunoglobulin domain and the DI–DII linker. The fusion peptide opens the trimer head, separating intersubunit contacts between DI and DII observed in the postfusion form²⁰. Fusion peptide residues 107–114 cross the DII G strand, whereas residues 115–117 form a β -strand with residues 370–373 (DII A-strand). The fusion peptide folds back on itself, forming a small hydrophobic core between its N- and C-terminal ends and making less extensive contacts with DIII. Proteolytic cleavage of F₀ might enable the N terminus of the fusion peptide to make additional contacts with DII and to affect intersubunit interactions.

Comparison of the SV5 and hPIV3 F structures

The SV5 F-GCNt and hPIV3 solF0 structures are in strikingly

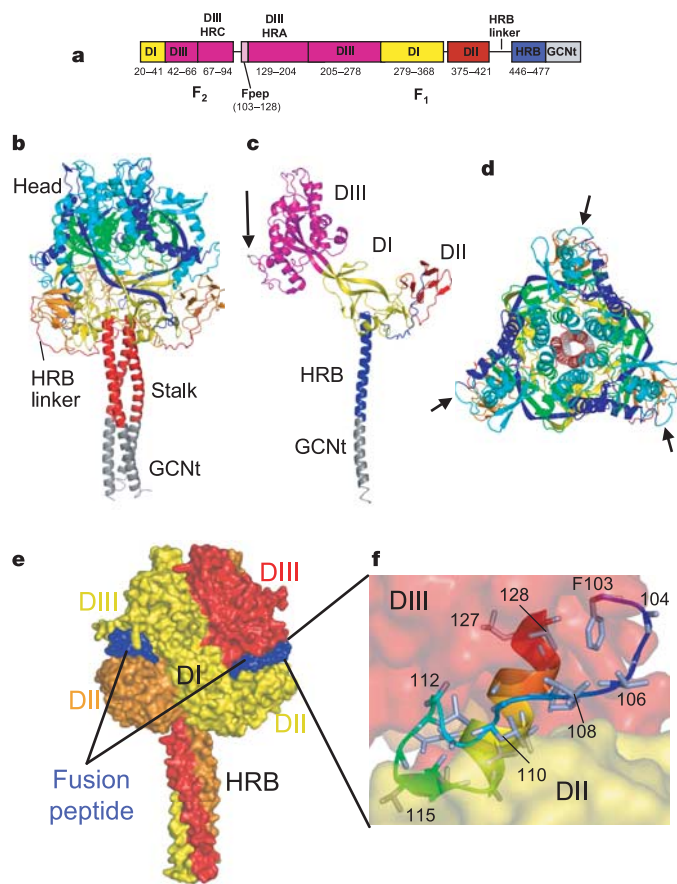


Figure 1 | Structure of SV5 F-GCNt. **a**, The F-GCNt domains. Important domains are highlighted in different colours and their corresponding residue ranges are indicated. **b**, Ribbon diagram of the F trimer, with each chain coloured by residue number in a gradient from blue (N terminus) to red (C terminus). The head and stalk regions are indicated. HRB linker residues (429–432) could not be modelled in one subunit and had high temperature factors in the other two. **c**, Ribbon diagram of one subunit of the F trimer, coloured by domain. The domains are labelled and the colours correspond to those in **a**. Arrow indicates the cleavage/activation site. **d**, Top view of the trimer, coloured as in **d**. Arrows indicate the cleavage/activation sites. **e**, Surface representation of the F trimer, coloured by subunit. The fusion peptide exposed surface is shown in blue. **f**, Close-up view of the fusion peptide (residues 103–128). The peptide is folded back on itself with a small hydrophobic core and contains a mixture of an extended chain, a β -strand and a C-terminal α -helix. The fusion peptide is sandwiched between two subunits of the trimer, between DII and DIII domains.

Table 1 | X-ray refinement statistics

Resolution (Å)	2.85
$R_{\text{work}}/R_{\text{free}}$	22.16 / 26.11
Number of atoms	
Protein	10,805
Water	130
B-factors	
Protein	65.4
Water	50.7
r.m.s. deviations	
Bond lengths (Å)	0.007
Bond angles (°)	1.42

different conformations (Fig. 2), consistent with a transition from pre- to postfusion forms. These conformations are related by flipping the stalk and transmembrane domains relative to the F head. Substantial compacting of the head is observed in hPIV3 solF0 as compared with SV5 F-GCNt. The DI domains pivot slightly inwards, shearing intersubunit contacts, and the DII domains swing across, contacting neighbouring subunits. Individual DI and DII domains in the two structures remain similar, superimposing with average root-mean-square (r.m.s.) deviations of $1.97 \text{ \AA}^2$ and $1.5 \text{ \AA}^2$, respectively. Potentially related forms of the RSV F protein have been observed in electron micrographs²⁹.

DIII undergoes major refolding between the two structures, projecting a new coiled coil (HRA) upwards and away from DI, the prefusion stalk and the viral membrane. The fusion peptide, located at the top of the HRA coiled coil, moves $\sim 115 \text{ \AA}$ from its initial position between subunits in the prefusion conformation, enabling DII to reposition. None of the postfusion HRA intersubunit coiled-coil contacts is observed in F-GCNt. Instead, they are replaced by two sets of six-helix rings at the DIII interfaces (Fig. 1d). For the HRA coiled coil to form, DIII must rotate and collapse inwards, further compacting the head.

In the prefusion conformation, HRA is broken up into four helices, two β -strands and five loop, kink or turn segments. Thus, the conformational changes in HRA involve the refolding of 11 distinct segments into a single, extended α -helical conformation (Fig. 3a). The conformational change also requires opening and translocation of the HRB stalk (Fig. 2). In the prefusion form, HRB is located at the base of the head region. During conversion to the

postfusion conformation, the HRB segments must separate and swing around the base of the head to pack against the HRA coiled coil.

HRA folds around the DIII core

The most marked changes in F occur around a relatively constant 'DIII core' that includes three antiparallel β -strands, the HRC α -helix, the C-terminal h4 helix of HRA and a helical bundle (Fig. 3b–f). In the prefusion conformation, HRA is folded around the DIII core. The globular form of the prefusion DIII suggests that the HRA conformation is trapped as monomeric subunits fold during F biosynthesis.

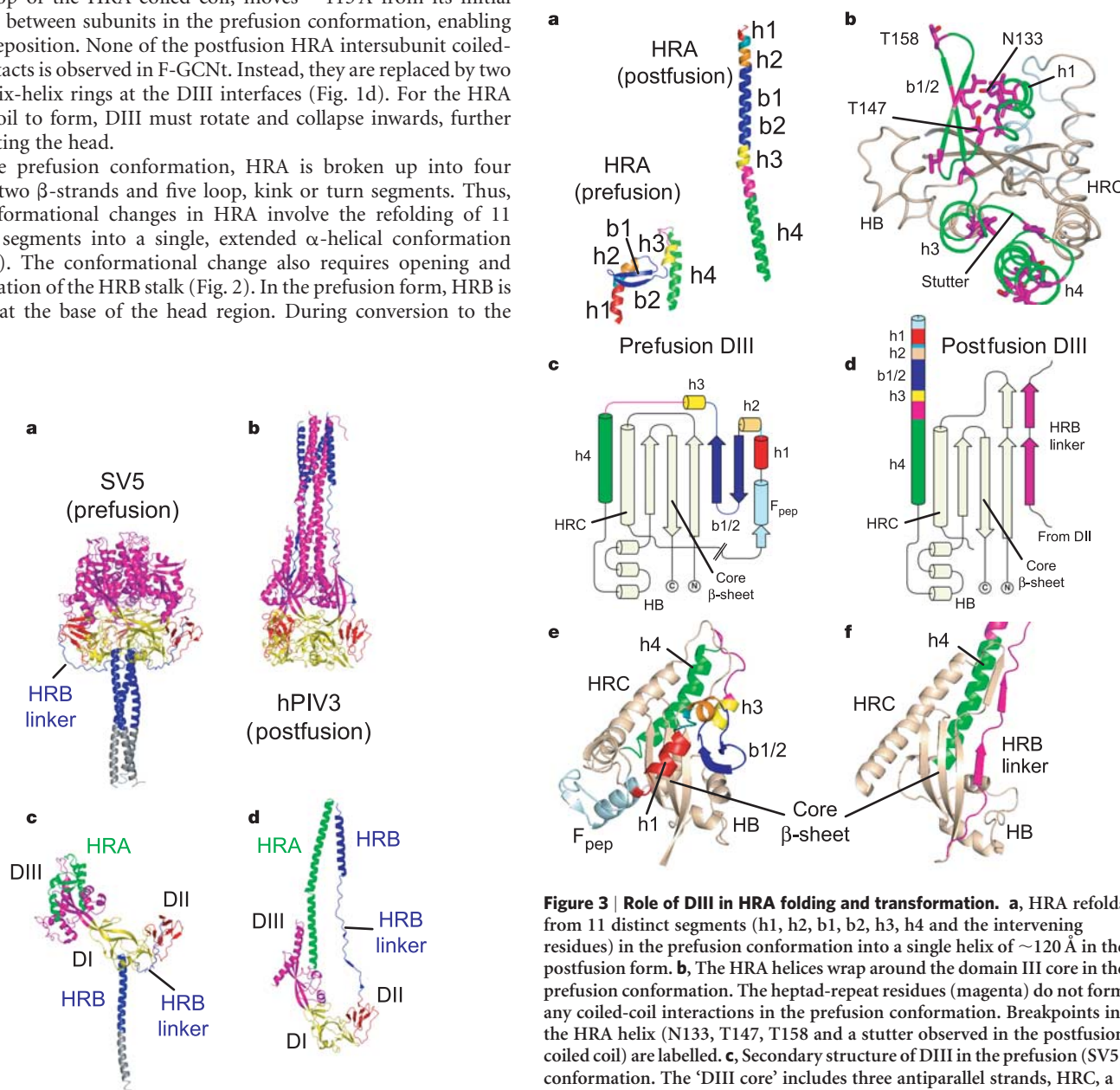


Figure 3 | Role of DIII in HRA folding and transformation. **a**, HRA refolds from 11 distinct segments (h1, h2, b1, b2, h3, h4 and the intervening residues) in the prefusion conformation into a single helix of $\sim 120 \text{ \AA}$ in the postfusion form. **b**, The HRA helices wrap around the domain III core in the prefusion conformation. The heptad-repeat residues (magenta) do not form any coiled-coil interactions in the prefusion conformation. Breakpoints in the HRA helix (N133, T147, T158 and a stutter observed in the postfusion coiled coil) are labelled. **c**, Secondary structure of DIII in the prefusion (SV5) conformation. The 'DIII core' includes three antiparallel strands, HRC, a helical bundle (HB) and h4 of HRA. Segments of HRA are coloured as in **a** and the cleavage site (//) and fusion peptide (F_{pep}) are indicated. The DIII core sheet is extended by the b1 and b2 strands from HRA. **d**, Secondary structure of DIII in the postfusion (hPIV3) conformation, coloured as in **c**. The DIII core sheet is extended by one strand from an HRB linker from a neighbouring subunit (magenta). **e**, Ribbon diagram of DIII in the prefusion conformation, coloured as in **c**. **f**, Ribbon diagram of DIII in the postfusion conformation, coloured as in **d**.

Figure 2 | Structural changes between the pre- and postfusion F protein conformations. **a**, Ribbon diagram of the SV5 F-GCNt trimer. DI is yellow, DII is red, DIII is magenta, HRB is blue and GCNt is grey. **b**, Ribbon diagram of the hPIV3 (postfusion) trimer, similarly oriented by DI and coloured as in **a**. **c**, Ribbon diagram of a single subunit of the SV5 F-GCNt trimer, coloured as in **a** except for residues of HRA, which are green. **d**, Ribbon diagram of a single subunit of the hPIV3 F trimer, coloured as in **c**.

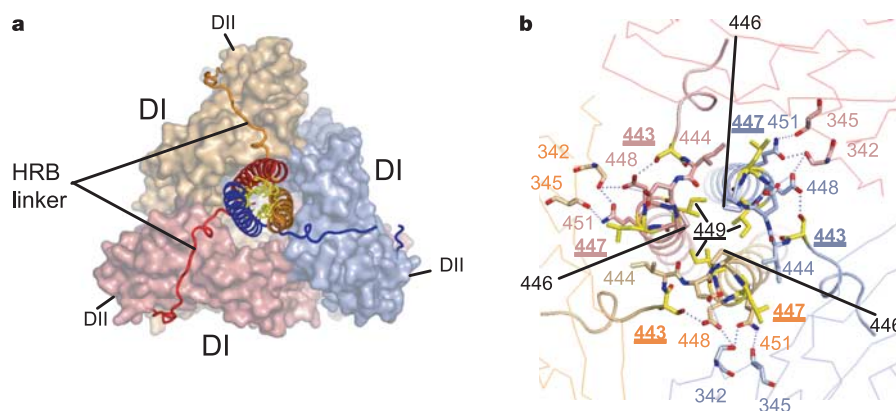


Figure 4 | HRB interacts with the base of the head region. **a**, View of HRB looking down the three-fold axis from the C terminus of the protein (transmembrane anchor/viral membrane view). The HRB stalk is aligned with the head region along the three-fold axis of the trimer. The HRB linkers extend from the end of DII around the outside of the trimer, before proceeding in towards the top of the HRB stalk, following open grooves in the DI domains. The HRB heptad-repeat residues form the core of the three-helix bundle in the prefusion state, but pack against the HRA coiled

In the prefusion form, HRA helical segments cover both sides of the DIII core β -strands (Fig. 3b), which also form a mixed parallel/antiparallel five-stranded sheet with the two HRA β -strands, b1 and b2 (Fig. 3c, e). In the postfusion form, the HRA b1 and b2 β -strands are replaced by a β -strand from the HRB linker of a neighbouring subunit. The HRB linker also forms an additional two-stranded β -sheet with DIII residues preceding HRC (Fig. 3d, f). Thus, folding of HRA onto the DIII core not only prevents formation of the HRA coiled coil, but also blocks this interaction of the HRB linker.

Hydrophobic heptad repeat (HR) residues form the core of the postfusion coiled coil, but they do not make these contacts in the prefusion conformation (Fig. 3b). Instead, some of the HR residues form a small hydrophobic core between the b1 and b2 β -strands and the h1 helix in the N-terminal region of the prefusion HRA. In the C-terminal region, the HR residues of h3 pack onto h4, whereas the HR residues of h4 interact with the DIII helical bundle and h3 from a neighbouring subunit. The h4 HR residues include many polar amino acids such as serine and threonine (Supplementary Fig. 7), which might facilitate packing changes between the pre- and postfusion forms.

The SV5 6HB structure¹⁷ contains two ion-binding sites in the HRA coiled-coil core and a 3-4-4-4-3 stutter near a hydrophobic pocket for HRB residues 447 and 449. In the prefusion structure, these features all map to breakpoints in the HRA helix (Fig. 3b), suggesting that the intrinsic instability of these regions in monomeric HRA may contribute to their role as conformational switch points. The ion-binding sites (N133 and T158) are located between helices h1 and h2 and strands b1 and b2, respectively (Fig. 3b). The HRA stutter residues are in the loop between helices h3 and h4; and T147, which forms stabilizing interactions with HRB in the 6HB, is in the loop between helix h2 and strand b1.

The head engages the top of the HRB stalk

The SV5 HRB region forms the helical stalk of the prefusion structure, along with the GCNt trimer, but isolated HRB peptide does not form stable three-helix bundles in aqueous solution¹⁹. In the F-GCNt structure, the three-fold axis of the HRB three-helix bundle is aligned along the three-fold axis of the head region, with only slight tilting of the stalk (Fig. 4a). The HRB linker segments (residues 422–445) adopt similar three-fold symmetric conformations, packing into open grooves on the outside of the trimer head.

Residues in HRB and the base of the head establish an interaction network between trimer subunits, appearing to position and to

coil in the postfusion form. F is coloured by subunit and the HR residues are shown as yellow sticks. **b**, View of the top of the HRB region looking down the three-fold axis from the base of the head. Interactions between subunits tie together the top of HRB and the base of the head region. The 'switch peptide' residues 443, 447 and 449 (yellow) are located here and form intersubunit interactions that explain their role in stabilizing the prefusion conformation.

nucleate the HRB helix (Fig. 4). S443 makes hydrogen bonds to D448, which makes hydrogen bonds to D445 of the same subunit and S342 of a neighbouring subunit. In the HRB helix, N451 interacts with S342 and T345 of the same neighbouring subunit. L447 packs into a hydrophobic pocket lined by I444, T357 and Q304, and residues 440–442 form a short two-strand parallel β -sheet with residues 358–359 in the same subunit. This conservation of symmetry and interactions between the head and stalk make it very unlikely that the presence of GCNt alters the native F structure.

The interactions at the base of the head and the stability of the HRB three-helix bundle probably regulate early steps in F activation. Mutations of residues 443, 447 and 449 destabilize the prefusion F conformation and have been predicted to form distinct interactions in the pre- and postfusion states^{30,31}. The change of S443 to proline could destabilize F by disrupting the hydrogen-bonding network described above (Fig. 4b). Mutations of L447 and I449 to aromatic residues also destabilize the prefusion conformation³¹, and these larger amino acids would disrupt hydrophobic contacts. Notably, L447 and I449 are in a helical conformation in the prefusion HRB stalk, but in an extended conformation in the postfusion 6HB. Other mutations affect SV5 F fusion activity^{30,32–34} and can be explained similarly by their locations in the prefusion form of F. Finally, the presence of the GCNt domain, or the natural transmembrane and cytoplasmic domains, undoubtedly stabilizes the HRB stalk, explaining why F-GCNt remains in the prefusion conformation, whereas secreted anchorless hPIV3 and NDV F proteins convert to the postfusion form^{20,21}.

Conclusions

The F-GCNt prefusion and the solF0 postfusion structures suggest how discrete refolding intermediates are coupled to the activation and progression of F-mediated membrane fusion. We propose the following model. In the first step, the HRB helices melt ('open-stalk form'; Fig. 5), breaking interactions at the base of the head, but leaving HRA in the prefusion conformation. This intermediate is consistent with the effects of mutating residues 443, 447 and 449, and with peptide inhibition data. HRA-derived peptides, which probably bind to the endogenous HRB segment, inhibit an early intermediate along the fusion pathway, whereas HRB-derived peptides inhibit a later intermediate by binding the endogenous HRA coiled coil¹⁶. Opening of the HRB stalk could initiate further changes in F by affecting the packing of DII and the fusion peptide (through the HRB linker), and by affecting the stability of the head intersubunit

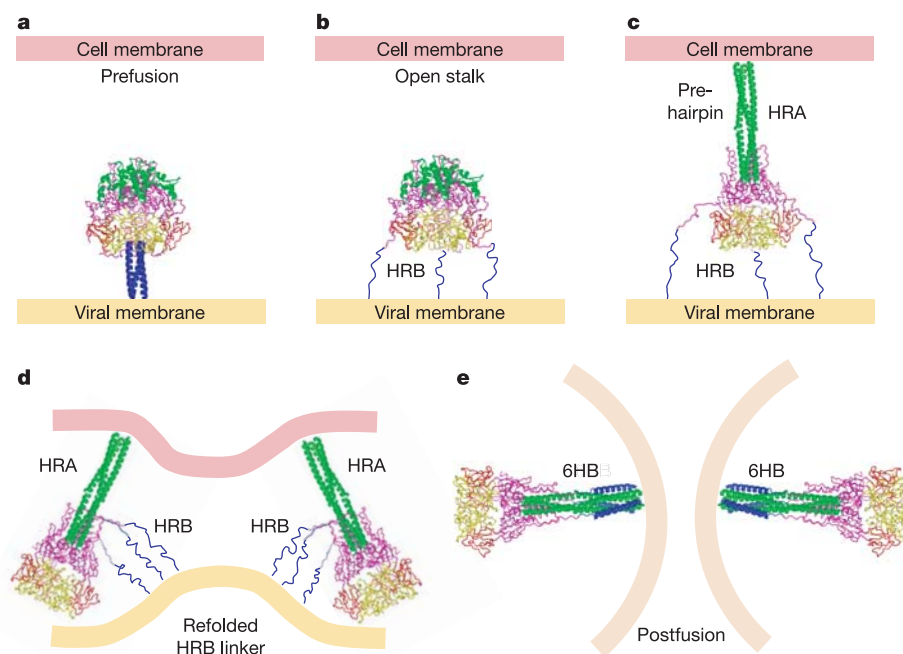


Figure 5 | Model of F-mediated membrane fusion. **a**, Structure of the prefusion conformation. HRB is blue, HRA is green, and DI, DII and DIII are yellow, red and magenta, respectively. **b**, 'Open stalk' conformation, in which the HRB stalk melts and separates from the prefusion head region. HRB is shown as three extended chains because the individual segments are unlikely to be helical. This conformation is consistent with a low-temperature intermediate that is inhibited by HRA peptides, but not HRB peptides. Mutations of the switch peptide residues 443, 447 and 449 would influence the formation of this intermediate by affecting stabilizing interactions between the prefusion stalk and head domains (see Fig. 4). **c**, A

pre-hairpin intermediate can form by refolding of DIII, facilitating formation of the HRA coiled coil and insertion of the fusion peptide into the target cell membrane. This intermediate can be inhibited by peptides derived from both HRA and HRB regions. **d**, Before formation of the final 6HB, folding of the HRB linker onto the newly exposed DIII core, with the formation of additional β -strands (see Fig. 3d, f), may stabilize the juxtaposition of viral and cellular membranes. **e**, The formation of the postfusion 6HB is tightly linked to membrane fusion and pore formation, juxtaposing the membrane-interacting fusion peptides and transmembrane domains.

contacts, which shift during the conformational transition. It seems possible that transient dissociation of the F trimer could occur, analogous to the dimer-to-trimer transition characterized in alpha-virus and flavivirus fusion proteins^{35,36}.

The open-stalk intermediate is probably then followed by refolding of DIII, assembly of the HRA coiled-coil and translocation of the fusion peptide towards the target cell membrane (Fig. 5). This pre-hairpin intermediate has been trapped and co-precipitated with HRB-peptides¹⁶. Removal of the fusion peptide from the intersubunit interfaces would enable an inward swing of DII and the formation of new contacts with DI of a neighbouring subunit, compacting the head. The refolding of DIII HRA would also expose its core β -sheet and, together with the inward movement of DII, enable the HRB linker (at the C terminus of DII) to form parallel β -strands with the DIII core, probably preceding and initiating the final positioning of HRB (Fig. 5). The assembly of the final 6HB completes the conformational change and membrane merger. Although proteolytic cleavage of the F protein is required for membrane fusion activity, it is apparently not required for formation of the postfusion conformation. The role of the HN or H protein in stimulating the F conformational change remains to be elucidated, but it could exert effects by, for example, influencing the stability of the F prefusion stalk in a receptor-dependent manner.

The F-GCNt structure shows how a metastable protein fold and its conformational transition to a more stable state can trigger membrane fusion. The folding of metastable proteins, as well as their activation, is not well understood but is a hallmark of class I viral fusion proteins such as F, influenza virus HA and HIV Env¹². The F structural changes are very different from those observed in influenza virus HA¹², which is the only other class I viral fusion protein for which we have both pre- and postfusion structures^{37,38}. However, new and potentially general concepts for these protein

machines emerge from a comparison of the F and HA fusion proteins: first, in the metastable state, the N-terminal (fusion-peptide-proximal) HRA segment is prevented from assembling into the postfusion coiled-coil structure; second, the fusion peptide is initially buried at subunit interfaces that undergo considerable reorganization between the pre- and postfusion states; third, refolding of HRA projects the fusion peptide away from the initial positions of the transmembrane anchors and viral membrane; and last, the C-terminal HRB region is prevented from adopting its final state both by the absence of the HRA coiled coil and by other inhibitory structural elements present in the prefusion conformation. For HIV Env, it is tempting to speculate that the gp120 'inner' domain, which switches conformation between the free and receptor-bound states, could act similarly to the F DIII core by regulating the folding of the HRA segment of HIV gp41 (refs 39, 40). The strategy of stabilizing the prefusion F conformation with GCNt may prove to be important for the elucidation of other viral fusion protein mechanisms.

METHODS

F protein expression and purification. The basic approaches for cloning, expression and purification of SV5 F-GCNt have been described for hPIV3 F (ref. 20). In brief, complementary DNA encoding a form of the SV5 (W3A strain) F protein (FR3) in which the furin cleavage site had been mutated to prevent intracellular processing⁴¹ was cloned into pMelBac (Invitrogen) by standard PCR protocols. A soluble form of F was generated that contained the honeybee melittin signal sequence in place of the F signal sequence and, at the C terminus, an isoleucine zipper domain (GCNt)^{23,42} in heptad repeat phase with HRB, followed by a factor Xa cleavage site and a His₆ tag. The nucleotide sequence of the construct was obtained with a 3100-Avant sequencer (Applied Biosystems). Recombinant baculovirus was generated with a Bac-N-Blue transfection kit (Invitrogen). The secreted F-GCNt protein was purified by Co²⁺-affinity chromatography.

Crystallization, structure determination and refinement. Crystals were grown

at 20 °C by the hanging-drop, vapour-diffusion method by mixing equal volumes of protein solution (10 mg ml⁻¹) and precipitant (0.9 M sodium potassium tartrate and 0.1 M imidazole; pH 8.0). Native and all heavy-atom-soaked crystals were transferred to cryosolvents consisting of 0.9 M sodium potassium tartrate, 0.1 M Tris-HCl or 0.1 M imidazole, and 20% propylene glycol (pH 8.0) for flash freezing in liquid nitrogen. All crystals belong to space group C222₁. Native data set 1 (in Tris-HCl) was collected to 3.0 Å and native data set 2 (in imidazole buffer) was collected to 2.85 Å (Supplementary Table 2). Crystals soaked with heavy atoms (10 mM OsCl₃, 1 mM PiP (di-μ-iodobis(ethylene-diamine)diplatinum(II) nitrate) or 20 mM AuCN₂) were collected to 4.5 Å (Supplementary Table 3). Diffraction data were processed with MOSFLM⁴³ and scaled and reduced to structure factor amplitudes by using the CCP4 suite of programs⁴³.

Heavy-atom positions were determined with the program SOLVE⁴⁴ and refinement, calculation of phases and density modification were carried out with SHARP⁴⁵ and Solomon⁴⁶. Using three platinum sites, four gold sites and three osmium sites, we obtained an initial 4.5 Å density-modified map and associated phases. These were used to position individual domains (DI–DIII) and a three-helix bundle from the hPIV3 F structure by using the six-dimensional phased translation search implemented in BRUTEPTF⁴⁷. The search resulted in good solutions for parts of each of DI–DIII and a three-helix bundle²⁰ that matched the observed electron density and could also be arranged as a three-fold symmetric trimer. The original experimental phases after solvent flattening were used, along with a partial molecular envelope and subunit NCS transformations based on this preliminary trimer model, as input to the program DM⁴³. Density modification and three-fold NCS averaging were carried out with phase extension from 8 to 3.0 Å resolution in 1,500 steps, yielding a readily interpretable electron density map.

The SV5 F-GCNt structure was refined by the program CNS⁴⁸, followed by manual rebuilding with the program O⁴⁹. After several rounds of manual rebuilding and refinement, the 2.85-Å native data set 2 was used to complete the model rebuilding and refinement after transferring the original *R*_{free} set. To improve the electron density for difficult regions, model phases were input into SHARP to improve the heavy-atom model and refinement. Additional heavy-atom sites were identified (Supplementary Table 3), and the resulting SHARP/Solomon electron density maps were of higher quality, enabling the HRB linker and C-terminal GCNt residues to be traced. The final structure has an *R*_{work} of 22.2% and an *R*_{free} of 26.1%. A Ramachandran plot shows that 99.3% of the residues are in the most favourable or additionally allowed regions. The final refinement statistics, native and heavy-atom data, and phasing statistics are summarized in Table 1 and Supplementary Tables 2 and 3. Figures were made with the programs Pymol (<http://pymol.sourceforge.net/>) and Topdraw⁴³.

Received 10 September; accepted 11 October 2005.

- Lamb, R. A. & Kolakofsky, D. in *Fields Virology* 4th edn (eds Knipe, D. M. & Howley, P. M.) 1305–1340 (Lippincott, Williams and Wilkins, Philadelphia, 2001).
- Murray, K. *et al.* A morbillivirus that caused fatal disease in horses and humans. *Science* **268**, 94–97 (1995).
- Chua, K. B. *et al.* Nipah virus: a recently emergent deadly paramyxovirus. *Science* **288**, 1432–1435 (2000).
- Dorig, R. E., Marciel, A., Chopra, A. & Richardson, C. D. The human CD46 molecule is a receptor for measles virus (Edmonston strain). *Cell* **75**, 295–305 (1993).
- Yanagi, Y., Ono, N., Tatsu, H., Hashimoto, K. & Minagawa, H. Measles virus receptor SLAM (CD150). *Virology* **299**, 155–161 (2002).
- Negrete, O. A. *et al.* EphrinB2 is the entry receptor for Nipah virus, an emergent deadly paramyxovirus. *Nature* **436**, 401–405 (2005).
- Bonaparte, M. I. *et al.* Ephrin-B2 ligand is a functional receptor for Hendra virus and Nipah virus. *Proc. Natl Acad. Sci. USA* **102**, 10652–10657 (2005).
- Feldman, S. A., Hendry, R. M. & Beeler, J. A. Identification of a linear heparin binding domain for human respiratory syncytial virus attachment glycoprotein G. *J. Virol.* **73**, 6610–6617 (1999).
- Lamb, R. A. Paramyxovirus fusion: a hypothesis for changes. *Virology* **197**, 1–11 (1993).
- Dutch, R. E., Jardetzky, T. S. & Lamb, R. A. Virus membrane fusion proteins: biological machines that undergo a metamorphosis. *Biosci. Rep.* **20**, 597–612 (2000).
- Earp, L. J., Delos, S. E., Park, H. E. & White, J. M. The many mechanisms of viral membrane fusion proteins. *Curr. Top. Microbiol. Immunol.* **285**, 25–66 (2005).
- Skehel, J. J. & Wiley, D. C. Receptor binding and membrane fusion in virus entry: the influenza haemagglutinin. *Annu. Rev. Biochem.* **69**, 531–569 (2000).
- Eckert, D. M. & Kim, P. S. Mechanisms of viral membrane fusion and its inhibition. *Annu. Rev. Biochem.* **70**, 777–810 (2001).
- Hernandez, L. D., Hoffman, L. R., Wolfsberg, T. G. & White, J. M. Virus–cell and cell–cell fusion. *Annu. Rev. Cell Dev. Biol.* **12**, 627–661 (1996).
- Furuta, R. A., Wild, C. T., Weng, Y. & Weiss, C. D. Capture of an early fusion-active conformation of HIV-1 gp41. *Nature Struct. Biol.* **5**, 276–279 (1998).
- Russell, C. J., Jardetzky, T. S. & Lamb, R. A. Membrane fusion machines of paramyxoviruses: capture of intermediates of fusion. *EMBO J.* **20**, 4024–4034 (2001).
- Baker, K., Dutch, R. E., Lamb, R. A. & Jardetzky, T. S. Structural basis for paramyxovirus-mediated membrane fusion. *Mol. Cell* **3**, 309–319 (1999).
- Melikian, G. B. *et al.* Evidence that the transition of HIV-1 gp41 into a six-helix bundle, not the bundle configuration, induces membrane fusion. *J. Cell Biol.* **151**, 413–423 (2000).
- Joshi, S. B., Dutch, R. E. & Lamb, R. A. A core trimer of the paramyxovirus fusion protein: parallels to influenza virus haemagglutinin and HIV-1 gp41. *Virology* **248**, 20–34 (1998).
- Yin, H. S., Paterson, R. G., Wen, X., Lamb, R. A. & Jardetzky, T. S. Structure of the uncleaved ectodomain of the paramyxovirus (hPIV3) fusion protein. *Proc. Natl Acad. Sci. USA* **102**, 9288–9293 (2005).
- Chen, L. *et al.* The structure of the fusion glycoprotein of Newcastle disease virus suggests a novel paradigm for the molecular mechanism of membrane fusion. *Structure* **9**, 255–266 (2001).
- Colman, P. M. & Lawrence, M. C. The structural biology of type I viral membrane fusion. *Nature Rev. Mol. Cell Biol.* **4**, 309–319 (2003).
- Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
- Waning, D. L., Russell, C. J., Jardetzky, T. S. & Lamb, R. A. Activation of a paramyxovirus fusion protein is modulated by inside-out signalling from the cytoplasmic tail. *Proc. Natl Acad. Sci. USA* **101**, 9217–9222 (2004).
- Stevens, J. Structure of the uncleaved human H1 haemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870 (2004).
- Chen, B. *et al.* A chimeric protein of simian immunodeficiency virus envelope glycoprotein gp140 and *Escherichia coli* aspartate transcarbamoylase. *J. Virol.* **78**, 4508–4516 (2004).
- Yang, X. *et al.* Modifications that stabilize human immunodeficiency virus envelope glycoprotein trimers in solution. *J. Virol.* **74**, 4746–4754 (2000).
- Yang, X. *et al.* Highly stable trimers formed by human immunodeficiency virus type 1 envelope glycoproteins fused with the trimeric motif of T4 bacteriophage fibrin. *J. Virol.* **76**, 4634–4642 (2002).
- Calder, L. J. *et al.* Electron microscopy of the human respiratory syncytial virus fusion protein and complexes that it forms with monoclonal antibodies. *Virology* **271**, 122–131 (2000).
- Paterson, R. G., Russell, C. J. & Lamb, R. A. Fusion protein of the paramyxovirus SV5: destabilizing and stabilizing mutants of fusion activation. *Virology* **270**, 17–30 (2000).
- Russell, C. J., Kantor, K. L., Jardetzky, T. S. & Lamb, R. A. A dual-functional paramyxovirus F protein regulatory switch segment: activation and membrane fusion. *J. Cell Biol.* **163**, 363–374 (2003).
- Russell, C. J., Jardetzky, T. S. & Lamb, R. A. Conserved glycine residues in the fusion peptide of the paramyxovirus fusion protein regulate activation of the native state. *J. Virol.* **78**, 13727–13742 (2004).
- Ito, M., Nishio, M., Komada, H., Ito, Y. & Tsurudome, M. An amino acid in the heptad repeat 1 domain is important for the haemagglutinin–neuraminidase-independent fusing activity of simian virus 5 fusion protein. *J. Gen. Virol.* **81**, 719–727 (2000).
- Tsurudome, M. *et al.* Hemagglutinin–neuraminidase-independent fusion activity of simian virus 5 fusion (F) protein: difference in conformation between fusogenic and nonfusogenic F proteins on the cell surface. *J. Virol.* **75**, 8999–9009 (2001).
- Gibbons, D. L. *et al.* Conformational change and protein–protein interactions of the fusion protein of Semliki Forest virus. *Nature* **427**, 320–325 (2004).
- Modis, Y., Ogata, S., Clements, D. & Harrison, S. C. Structure of the dengue virus envelope protein after membrane fusion. *Nature* **427**, 313–319 (2004).
- Wilson, I. A., Skehel, J. J. & Wiley, D. C. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* **289**, 366–375 (1981).
- Bullough, P. A., Hughson, F. M., Skehel, J. J. & Wiley, D. C. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* **371**, 37–43 (1994).
- Chen, B. *et al.* Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* **433**, 834–841 (2005).
- Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**, 648–659 (1998).
- Paterson, R. G., Shaughnessy, M. A. & Lamb, R. A. Analysis of the relationship between cleavability of a paramyxovirus fusion protein and length of the connecting peptide. *J. Virol.* **63**, 1293–1301 (1989).
- Harbury, P. B., Kim, P. S. & Alber, T. Crystal structure of an isoleucine-zipper trimer. *Nature* **371**, 80–83 (1994).
- Collaborative Computational Project No. 4, The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).

45. de La Fortelle, E. & Bricogne, G. in *Methods in Enzymology, Macromolecular Crystallography* (eds Sweet, R. M. & Carter, C. W.) 472–494 (Academic, New York, 1997).
46. Abrahams, J. P. & Leslie, A. G. Methods used in the structure determination of bovine mitochondrial F₁ ATPase. *Acta Crystallogr. D* **52**, 30–42 (1996).
47. Strokopytov, B. V. *et al.* Phased translation function revisited: structure solution of the cofilin-homology domain from yeast actin-binding protein 1 using six-dimensional searches. *Acta Crystallogr. D* **61**, 285–293 (2005).
48. Brünger, A. T. *X-PLOR: A System for X-ray Crystallography and NMR* (Yale Univ., New Haven, 1992).
49. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Shaughnessy Nagel for technical assistance

and beamline staff for assistance. Data were collected at the DND-CAT and LS-CAT beamlines at the Advanced Photon Source and at the Howard Hughes Medical Institute (HHMI) beamlines at the Advanced Light Source. Support for the Northwestern Center for Structural Biology from the R. H. Lurie Cancer Center is acknowledged. This research was supported in part by NIH research grants (to T.S.J. and R.A.L.). H.-S.Y. is an Associate and R.A.L. is an investigator of the HHMI, and T.S.J. is a Scholar of the Leukemia and Lymphoma Society of America.

Author Contributions H.-S.Y., X.W., R.G.P., R.A.L. and T.S.J. performed the experimental work. H.-S.Y. and T.S.J. performed computational analysis of the data. H.-S.Y., R.G.P., R.A.L. and T.S.J. interpreted the results and H.-S.Y., X.W., R.G.P., R.A.L. and T.S.J. wrote the manuscript.

Author Information Coordinates and structure factor amplitudes have been deposited in the Protein Data Bank (PDB ID code 2B9B). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.S.J. (tedj@northwestern.edu).

Radioactive ^{26}Al from massive stars in the Galaxy

Roland Diehl¹, Hubert Halloin¹, Karsten Kretschmer¹, Giselher G. Lichti¹, Volker Schönfelder¹, Andrew W. Strong¹, Andreas von Kienlin¹, Wei Wang¹, Pierre Jean², Jürgen Knödseder², Jean-Pierre Roques², Georg Weidenspointner², Stephane Schanne³, Dieter H. Hartmann⁴, Christoph Winkler⁵ & Cornelia Wunderer⁶

Gamma-rays from radioactive ^{26}Al (half-life $\sim 7.2 \times 10^5$ years) provide a 'snapshot' view of continuing nucleosynthesis in the Galaxy¹. The Galaxy is relatively transparent to such γ -rays, and emission has been found concentrated along its plane². This led to the conclusion¹ that massive stars throughout the Galaxy dominate the production of ^{26}Al . On the other hand, meteoritic data show evidence for locally produced ^{26}Al , perhaps from spallation reactions in the protosolar disk^{3–5}. Furthermore, prominent γ -ray emission from the Cygnus region suggests that a substantial fraction of Galactic ^{26}Al could originate in localized star-forming regions. Here we report high spectral resolution measurements of ^{26}Al emission at 1808.65 keV, which demonstrate that the ^{26}Al source regions corotate with the Galaxy, supporting its Galaxy-wide origin. We determine a present-day equilibrium mass of $2.8 (\pm 0.8)$ solar masses of ^{26}Al . We use this to determine that the frequency of core collapse (that is, type Ib/c and type II) supernovae is $1.9 (\pm 1.1)$ events per century.

Excess ^{26}Mg found in meteorites shows that the hot disk-accretion phase of the presolar nebula was apparently characterized^{3,4} by an amount of radioactive ^{26}Al (relative to the stable ^{27}Al isotope) with a rather well-determined $^{26}\text{Al}/^{27}\text{Al}$ ratio of $\sim 4.5 \times 10^{-5}$. This is surprising, given that ^{26}Al decays within ~ 1 Myr: the time it takes for a parental molecular cloud to form protostellar disks after decoupling from nucleosynthetically enriched interstellar gas is much longer⁶. Therefore, the meteoritic determinations of the $^{26}\text{Al}/^{27}\text{Al}$ ratio have been interpreted as an *in situ* ^{26}Al enrichment of the young solar nebula⁷, either by a nearby supernova or by an AGB star event injecting fresh nucleosynthesis products at the 'last moment', or by enhanced cosmic-ray nucleosynthesis in the magnetically active early Sun with its accretion disk⁵. The mean ^{26}Al content of the interstellar medium in the Galaxy would therefore decouple from the solar value.

Observation of 1808.65-keV γ -rays from the decay of radioactive ^{26}Al in the interstellar medium, however, demonstrated that ^{26}Al nucleosynthesis does occur in the present Galaxy. The irregular distribution^{2,8} of ^{26}Al emission seen along the plane of the Galaxy provided the main argument for the idea that massive stars dominate the production of ^{26}Al (ref. 1). Massive stars preferentially form in clusters; some of the nearby massive-star regions appear prominent in ^{26}Al emission (for example, in the Cygnus region), while others do not. Because the massive star census in the Galaxy is well known only out to distances of a few kiloparsecs, and many regions of the Galaxy are occulted for direct measurements, we are left with considerable uncertainty about a Galaxy-wide interpretation of the γ -ray measurements, with respect to the possibility of localized efficient ^{26}Al -producing regions. The total amount of ^{26}Al in the Galaxy, and hence the mean interstellar $^{26}\text{Al}/^{27}\text{Al}$ ratio, is thus rather uncertain.

If ^{26}Al sources are indeed distributed throughout the Galaxy,

Galactic rotation will cause Doppler shifts of the γ -ray line energy, depending on the location of the source region within the Galaxy. Offsets in the line energy range up to 0.25 keV, and are particularly pronounced towards longitudes around $\pm 30^\circ$ (ref. 9). From 1.5 years of data from our Ge spectrometer telescope (SPI) on the INTEGRAL γ -ray observatory of the European Space Agency (launched in October 2002) we derived new spectra of celestial ^{26}Al emission (see Supplementary Information for details of observations and analysis method). From the inner Galaxy, we obtain an ^{26}Al measurement at 16σ above the background, and significantly ($>3\sigma$) detect the ^{26}Al line emission in six 0.5-keV-wide energy bins across the line centre—this allows an unprecedented investigation of ^{26}Al γ -ray line

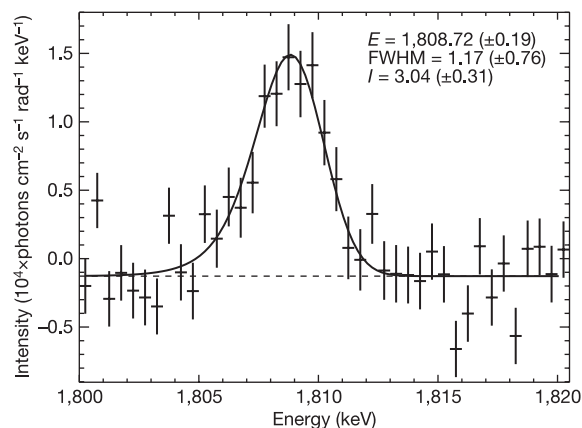


Figure 1 | Measurement of the ^{26}Al line from the inner Galaxy region with SPI/INTEGRAL. Our INTEGRAL sky exposure is fairly symmetric around the centre of the galaxy and extends over the full Galactic plane, though emphasizing the range of $\pm 45^\circ$ longitude and $\pm 15^\circ$ latitude (see Supplementary Information). This spectrum (shown with error bars of s.d.) was derived from sky model fitting to the set of 19 Ge detector count spectra for each of the 7,130 spacecraft pointings, using the COMPTEL ^{26}Al image as a model for the spatial distribution of emission, and was background-modelled from auxiliary measurements. The fit (solid line) combines the instrumental resolution, which is derived by accounting for degradation from cosmic-ray degradation irradiation and for annealings during the time of our measurement, with a gaussian for the intrinsic, astrophysical ^{26}Al linewidth. For this integrated result from the inner Galaxy, the line centre is determined to be at $1808.72 (\pm 0.2_{\text{stat}} \pm 0.1_{\text{syst}})$ keV (statistical and systematic uncertainties shown), well within the laboratory value for the ^{26}Al line of $1808.65 (7)$ keV. The integrated intensity from the inner region of the Galaxy is determined to be $3.3 (\pm 0.4) \times 10^{-4}$ photons $\text{cm}^{-2} \text{s}^{-1} \text{rad}^{-1}$, averaging over this and other plausible spatial-distribution models (hence uncertainty shown as sum of statistical and systematic; see Supplementary Information for method and details).

¹Max-Planck-Institut für extraterrestrische Physik, D-85748 Garching, Germany. ²Centre d'Etude Spatiale des Rayonnements and Université Paul Sabatier, 31028 Toulouse, France. ³DSM/DAPNIA/Service d'Astrophysique, CEA Saclay, 91191 Gif-Sur-Yvette, France. ⁴Clemson University, Clemson, South Carolina 29634-0978, USA. ⁵ESA/ESTEC, SCI-SD 2201 AZ Noordwijk, The Netherlands. ⁶Space Sciences Laboratory, Berkeley, California 94720, USA.

parameters (Fig. 1). We find that the observed lineshape approaches the one expected from instrumental resolution. The linewidth of celestial emission, which would be observable as additional broadening, is therefore rather small. The high spectral precision of our instrument can now be combined with its imaging capability to search for signatures of Galactic rotation, that is, to determine ^{26}Al line-centre energies for different regions along the Galactic plane.

Our method of deriving spectra must adopt a model for the distribution of emission over the sky (see Supplementary Information). When we split such a model in the region of interest ($-40^\circ \dots +40^\circ$) into three longitude segments through cuts at -10° and 10° , and determine ^{26}Al line energies for these regions, we systematically find that the ^{26}Al line energy clearly falls above the laboratory energy east of the Galactic centre, and slightly below towards the west (see Fig. 2 and Supplementary Information). Such a signature is expected from Galactic rotation. This strongly supports the view that the observed ^{26}Al source regions are located in the inner region of the Galaxy, rather than in localized foreground regions, as they appear to follow the global Galactic kinematics. This affirms a Galaxy-wide interpretation of the ^{26}Al γ -ray measurement, which had previously been argued for indirectly on the grounds of correlating the ^{26}Al γ -ray image with different tracers of candidate sources^{1,10}.

The total ^{26}Al γ -ray flux that we obtain is $3.3 (\pm 0.4) \times 10^{-4}$ photons $\text{cm}^{-2} \text{s}^{-1}$. This value is conventionally quoted for the inner Galaxy region ($-30^\circ < l < 30^\circ$; $-10^\circ < b < 10^\circ$) and the combined uncertainty from statistics and systematics is shown; the flux varies by $\sim 4\%$ when we use a range of models, which we consider to be plausible tracers of ^{26}Al sources^{1,10} (see Supplementary Information). This can be converted to an equilibrium mass produced by ongoing nucleosynthesis throughout the Galaxy in steady state, once the three-dimensional spatial distribution is known. Our best three-dimensional model is based on free electrons liberated by ionizing radiation from massive stars—such electrons can be measured at radio wavelengths, and the results have been translated into a geometrical model^{11,12}. Using this and alternative plausible models¹³ (see Supplementary Information), we infer a mass of

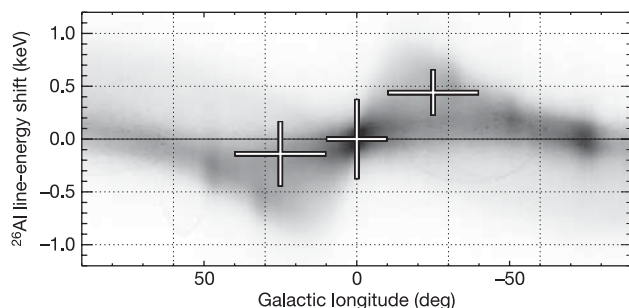


Figure 2 | Line position shifts with viewing directions along the inner Galaxy. Galactic rotation will shift the observed ^{26}Al line energy owing to the Doppler effect, to appear blue-shifted at negative longitudes and red-shifted at positive longitudes. Our expectations (greyscale) can be modelled from the Galactic rotation curve and a hypothetical three-dimensional distribution of ^{26}Al sources. From such models, we typically expect Doppler shifts of 0.25 keV, varying by ~ 0.05 keV with assumptions about inner-Galaxy rotation and spatial source distribution (see Supplementary Information), for the integrated longitude ranges $40^\circ \dots 10^\circ$ and $-10^\circ \dots -40^\circ$, respectively. We show here one of our models (in greyscale), based on free electrons in the interstellar medium^{11,12}, with an exponential distribution perpendicular to the Galactic plane (scale height 180 pc); for our longitude intervals $40^\circ \dots 10^\circ$ and $-10^\circ \dots -40^\circ$; this predicts integrated Doppler shifts of -0.22 keV and $+0.24$ keV, respectively. The three line energy measurements (s.d. error bars shown) extracted from our observations are consistent with the Galactic-rotation explanation at the 94% probability level (see Supplementary Information for method and details).

$2.8 M_\odot$ (where $M_\odot$ is the mass of the Sun) of ^{26}Al in the entire Galaxy. We estimate this value to be uncertain by $\pm 0.8 M_\odot$, from the combined statistical uncertainty of our measurements and the spatial-model uncertainty. The $^{26}\text{Al}/^{27}\text{Al}$ ratio implied^{13,14} for the average interstellar medium is 8.4×10^{-6} (assuming an interstellar gas mass of $4.95 \times 10^9 M_\odot$, and a numerical abundance of $\log N(^{27}\text{Al}) = 6.4$, on a scale given by $\log N(\text{H}) = 12$), about one order of magnitude lower than the solar-nebula value.

In conjunction with stellar yields and a distribution function of stellar birth masses this provides an independent estimate of the star-formation rate in the Galaxy. Although we know the star-formation rates for external galaxies and specific regions in the solar neighbourhood reasonably well, the star-formation rate of the Galaxy as a whole is much less certain, because of occultation from interstellar clouds or other biases. Values based on optical-to-infrared tracers range from 0.8 to $13 M_\odot$ per year^{15–18} (see Supplementary Information). The γ -ray technique has the advantage in that it measures the rate in penetrating radiation over the full Galaxy, and averages over a timescale associated more closely with one (current) generation of massive stars (1 Myr). The signature of Galactic rotation in the ^{26}Al γ -ray line reaffirms that large-scale distributions for tracers of ^{26}Al sources can be applied to obtain a census of massive stars in the Galaxy.

Theoretical nucleosynthesis yields have been derived for massive stars, which are presumed to dominate the Galactic ^{26}Al budget, specifically for core-collapse supernovae and for the preceding Wolf-Rayet phases. Such models have been shown to match the chemical history of the Galaxy as reflected in the abundances of chemical elements to within a factor of two¹⁹, which is an impressive success. ^{26}Al yields (wind-phase and explosive yields) from recent models of several independent research groups (Woosley, S. E., Heger, A. & Hoffman, R. D., manuscript in preparation; Limongi, M., & Chieffi, A., personal communication; refs 20 and 21) converge within about 50% over the full mass range (see Supplementary Information). Yields are moderated by the steep initial mass function (IMF), $\xi \approx m^{-\alpha}$, in our relevant mass range ~ 10 – $120 M_\odot$. We use the Salpeter IMF ($\xi \approx m^{-2.7}$) for this higher-mass range, supported by a wide range of astronomical constraints²², to obtain an average ejected mass per massive star of ^{26}Al of $Y_{26} = 1.4 \times 10^{-4} M_\odot$. With our measured amount of ^{26}Al , this IMF-weighted ^{26}Al yield implies a rate of core-collapse supernovae in the Galaxy, averaged over the radioactive lifetime of ^{26}Al (1.04 Myr); this represents the current core-collapse supernova rate: Evolutionary times for star clusters are ~ 10 – 100 Myr, while the formation times of star clusters from giant molecular clouds are ~ 100 Myr, and the Galaxy's age is ~ 12 Gyr. (See Supplementary Information for the method of calculation and discussion of uncertainties). Our ^{26}Al measure implies a rate of (1.9 ± 1.1) core collapses per century, corresponding to a star-formation rate of $\sim 4 M_\odot \text{yr}^{-1}$, or a stellar production rate of ~ 7.5 stars per year. Our high-resolution spectroscopy of ^{26}Al with SPI/INTEGRAL data shows that the Galaxy produces stars at a moderate rate, typical for spiral galaxies of similar type and luminosity.

Received 28 June; accepted 19 October 2005.

1. Prantzos, N. & Diehl, R. Radioactive ^{26}Al in the Galaxy: observations versus theory. *Phys. Rep.* **267**(1), 1–69 (1996).
2. Diehl, R. et al. COMPTEL observations of Galactic ^{26}Al . *Astron. Astrophys.* **298**, 445–460 (1995).
3. MacPherson, G. J., Davis, A. M. & Zinner, E. K. Distribution of ^{26}Al in the early solar system. *Meteoritics* **30**, 365–386 (1995).
4. Young, E. D. et al. Supra-canonical $^{26}\text{Al}/^{27}\text{Al}$ and the residence time of CAs in the solar protoplanetary disk. *Science* **308**, 223–227 (2005).
5. Lee, T. et al. Protostellar cosmic rays and extinct radioactivities in meteorites. *Astrophys. J.* **506**, 898–912 (1998).
6. Ward-Thompson, D. Isolated star formation: from cloud formation to core collapse. *Science* **295**, 76–81 (2002).
7. Meyer, B. S. & Clayton, D. D. Short-lived radioactivities and the birth of the Sun. *Space Sci. Rev.* **92**, 133–152 (2000).

8. Plüschke, S., et al. in *Exploring the Gamma-Ray Universe (4th INTEGRAL Workshop)* (eds Gimenez, A., Reglero, V. & Winkler, C.) 55–58 (ESA Special Publication, ESA-SP 459, Noordwijk, 2001).
9. Kretschmer, K., Diehl, R. & Hartmann, D. H. Line shape diagnostics of Galactic ^{26}Al . *Astron. Astrophys.* **412**, L77–L81 (2003).
10. Knödseder, J. et al. A multiwavelength comparison of COMPTEL ^{26}Al line data. *Astron. Astrophys.* **344**, 68–82 (1999).
11. Taylor, J. H. & Cordes, J. M. Pulsar distances and the Galactic distribution of free electrons. *Astrophys. J.* **411**, 674–684 (1993).
12. Cordes, J. M. & Lazio, T. J. W. NE2001. I. A new model for the Galactic distribution of free electrons and its variability. Preprint at (<http://arXiv.org/astro-ph/0207156>) (2002).
13. Robin, A. C., Reylé, C., Derriere, R. & Picaud, S. A synthetic view on structure and evolution of the Milky Way. *Astron. Astrophys.* **409**, 523–540 (2003).
14. Asplund, M., Grevesse, N. & Sauval, A. J. in *Cosmic Abundances as Records of Stellar Evolution and Nucleosynthesis in Honor of David L. Lambert* Vol. 336 25–39 (ASP Conf. Ser., 2005).
15. Gilmore, G. The star formation history of the Milky Way. *ASP Conf. Proc.* **230**, 3–12 (2001).
16. Boissier, S. & Prantzos, N. Chemo-spectrophotometric evolution of spiral galaxies—I. The model and the Milky Way. *Mon. Not. R. Astron. Soc.* **307**, 857–876 (1999).
17. McKee, C. F. & Williams, J. P. The luminosity function of OB associations in the Galaxy. *Astrophys. J.* **476**, 144–165 (1997).
18. Reed, C. B. New estimates of the solar-neighborhood massive star birthrate and the Galactic supernova rate. *Astron. J.* **130**, 1652–1657 (2005).
19. Rauscher, T., Heger, A., Hoffman, R. D. & Woosley, S. E. Nucleosynthesis in massive stars with improved nuclear and stellar physics. *Astrophys. J.* **576**, 323–348 (2002).
20. Limongi, M. & Chieffi, A. ^{26}Al and ^{60}Fe from massive stars. *Nucl. Phys. A* **758**, 11c–14c (2005).
21. Palacios, A. et al. New estimates of the contribution of Wolf Rayet stellar winds to the Galactic ^{26}Al . *Astron. Astrophys.* **429**, 613–624 (2005).
22. Kroupa, P. The initial mass function of stars. *Science* **295**, 82–91 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This paper is based on observations with INTEGRAL, an ESA project with instruments and a science data center funded by ESA member states (especially the PI countries: Denmark, France, Germany, Italy, Switzerland, Spain), Czech Republic and Poland, and with the participation of Russia and the USA. The SPI project has been completed under the responsibility and leadership of CNES/France. The SPI anticoincidence system is supported by the German government. We are grateful to ASI, CEA, CNES, DLR, ESA, INTA, NASA and OSTC for support. We are grateful to Alessandro Chieffi, Nikos Prantzos, and Stan Woosley for discussions of theoretical nucleosynthesis yields.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.D. (rod@mpe.mpg.de).

LETTERS

Charon's radius and atmospheric constraints from observations of a stellar occultation

A. A. S. Gulbis¹, J. L. Elliot^{1,2,3}, M. J. Person¹, E. R. Adams¹, B. A. Babcock^{4,6}, M. Emilio⁷, J. W. Gangestad^{5,6}, S. D. Kern¹, E. A. Kramer¹, D. J. Osip⁸, J. M. Pasachoff⁵, S. P. Souza⁵ & T. Tuvikene^{6,9}

The physical characteristics of Pluto and its moon, Charon, provide insight into the evolution of the outer Solar System. Although previous measurements have constrained the masses of these bodies^{1,2}, their radii and densities have remained uncertain. The observation of a stellar occultation by Charon in 1980 established a lower limit on its radius of 600 km (ref. 3) (later refined to 601.5 km; ref. 4) and suggested a possible atmosphere⁴. Subsequent, mutual event modelling yielded a range of 600–650 km (ref. 5), corresponding to a density of $1.56 \pm 0.22 \text{ g cm}^{-3}$ (refs 2, 5). Here we report multiple-station observations of a stellar occultation by Charon. From these data, we find a mean radius of $606 \pm 8 \text{ km}$, a bulk density of $1.72 \pm 0.15 \text{ g cm}^{-3}$, and rock-mass fraction 0.63 ± 0.05 . We do not detect a significant atmosphere and place 3σ upper limits on atmospheric number densities for candidate gases. These results seem to be consistent with collisional formation for the Pluto–Charon system in which the precursor objects may have been differentiated⁶, and they leave open the possibility of atmospheric retention by the largest objects in the outer Solar System.

Observing light from a star as an object passes in front of it—a stellar occultation—produces data with the highest spatial resolution available from Earth-based observing methods. Such data are critical for determining accurate sizes and probing atmospheres of distant Solar System objects; however, observations can be difficult because of the required geographic and temporal precision. A stellar occultation by Charon has been observed only once before and resulted in a single chord³.

An occultation by Charon of the star ‘C313.2’ (originally identified as a Pluto occultation candidate star ‘P313.2’⁷; UCAC2 26257135, with R- and K-band magnitudes respectively $R = 14.8$ and $K = 12.2$) was predicted to occur on 11 July 2005 (UT) (<http://occult.mit.edu/research/occultations/Charon/C313.2.html>). We arranged to observe the event from four sites in South America using five telescopes. The telescopes were the 0.6-m at Laboratório Nacional de Astrofísica's Observatório Pico dos Dias, Itajubá, Brazil; the 0.84-m at Observatório Cerro Armazones, Antofagasta, Chile; the 2.5-m du Pont and 6.5-m Clay at Las Campanas Observatory, La Serena, Chile; and the 8-m Gemini South on Cerro Pachón, La Serena, Chile. Data were successfully obtained at all but Pico dos Dias, where the occultation was unobservable owing to clouds. The occultation occurred four days after new moon, and image quality ranged from excellent to poor at the successful observing stations (0.5–4 arcsec). The instruments employed for the observations, excluding Gemini South, were POETS (Portable Occultation, Eclipse, and Transit Systems). Each system consisted of a high-speed Andor Ixon camera, an instrument control computer, and a GPS receiver to trigger frames and establish

accurate timing. Gemini South observations were performed using the Acquisition Camera. Observation details for each site are provided in Supplementary Information.

Charon's aspect at the time of the occultation is shown in Fig. 1, overlaid with the paths of each observed chord. The stations were spread approximately 580 km perpendicular to Charon's motion and spanned just under half of the shadow width, with stations on both sides of the centreline. Astrometric data were recorded at each successful site roughly an hour before and after the event, with twenty minutes of contiguous high-speed data encompassing the predicted midtimes. For reducing the data that spanned the occultation, a range of square apertures was selected, containing C313.2, Pluto and Charon. We used the astrometric frames, in which C313.2

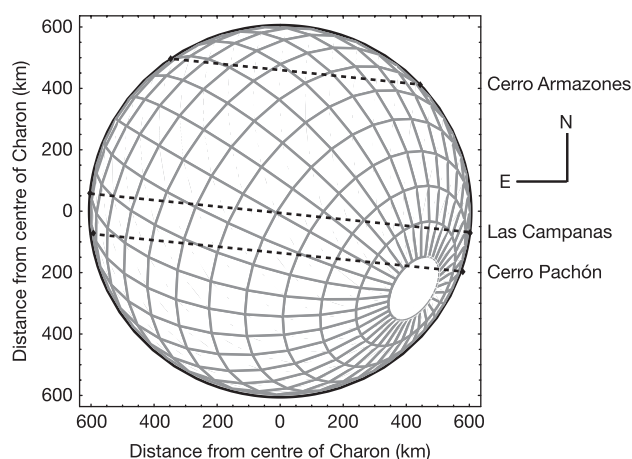


Figure 1 | Observed occultation chords. Overlaid on a diagram of Charon, dashed bold lines indicate the occultation chord paths as they appeared from each successful observing station. The coordinates (longitude, latitude) of each station are: Cerro Armazones ($70^{\circ}11'46'' \text{ W}$, $24^{\circ}35'52'' \text{ S}$), Las Campanas du Pont ($70^{\circ}42'13'' \text{ W}$, $29^{\circ}00'26'' \text{ S}$), Las Campanas Clay ($70^{\circ}41'33'' \text{ W}$, $29^{\circ}00'51'' \text{ S}$) and Cerro Pachón ($70^{\circ}43'24'' \text{ W}$, $30^{\circ}13'42'' \text{ S}$). Coordinates were obtained from POETS GPS surveys at all locations except Cerro Pachón, for which we reference the Astronomical Almanac²⁸. The du Pont and Clay telescopes at Las Campanas are geographically close enough that their chords are indistinguishable in this plot. Charon's south pole (IAU coordinate convention) is visible in the lower right. The distance to Charon at the time of the occultation was 30.07 AU, and the geocentric shadow velocity was 20.93 km s^{-1} . The geometric limb times during immersion and emersion for each of these chords were used to calculate Charon's radius.

¹Department of Earth, Atmospheric, and Planetary Sciences, ²Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA.

³Lowell Observatory, Flagstaff, Arizona 86001, USA. ⁴Physics Department, ⁵Astronomy Department, Williams College, Williamstown, Massachusetts 01267-2565, USA.

⁶Instituto de Astronomia, Universidad Católica del Norte, Avda. Angamos 0610, Antofagasta, Chile. ⁷Departamento de Geociências, Universidade Estadual de Ponta Grossa, Paraná, Brazil. ⁸Las Campanas Observatory, Carnegie Observatories, Casilla 601, La Serena, Chile. ⁹Vrije Universiteit Brussel, Brussels, Belgium.

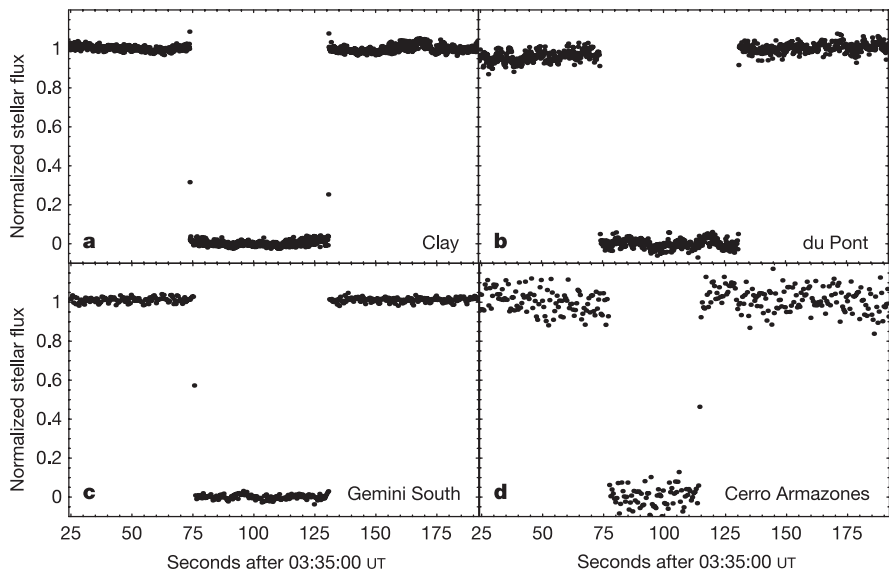


Figure 2 | Light curves showing the occultation of C313.2 by Charon. These plots display the normalized flux of the star observed at each of the four successful telescopes before, during and after the occultation. Telescopes and data rates for each of the plots are **a**, the Las Campanas 6.5-m Clay at 10 Hz, **b**, the Las Campanas 2.5-m du Pont at 5 Hz, **c**, the Cerro Pachón 8-m Gemini South at ~2.24 Hz, and **d**, the Cerro Armazones 0.84-m telescope at 2 Hz. Excluding Gemini South, data were obtained using POETS (frame transfer time of 1.74 ms per exposure) and no filter. Gemini South employed the Acquisition Camera and the R_G0154 filter (610–750 nm), with an integration time of 0.3 s and 0.146 ± 0.001 s required for readout. The data

points are plotted at the midpoint of each exposure, with respect to seconds after 03:35:00 UT. Geometric limb times from model fits to the light curves are **a**, immersion 73.744 ± 0.001 s, emersion 130.563 ± 0.002 s, **b**, immersion 73.792 ± 0.005 s, emersion 130.609 ± 0.004 s, **c**, immersion 75.50 ± 0.15 s, emersion 130.55 ± 0.15 s and **d**, immersion 76.99 ± 0.04 s, emersion 114.28 ± 0.03 s. Signal-to-noise ratios of the data, calculated from light-curve fit residuals and normalized to a 1 s cycle, are **a**, 273, **b**, 117, **c**, 108, and **d**, 28. The anomalously high points prior to immersion and post emersion in the Clay data result from the first diffraction fringe.

remained clearly resolved from Pluto–Charon, to measure the spatial offset between a standard star and C313.2. The centre of each aperture was then set by determining the centre of the comparison star and employing the offset. The sum of the signal in the aperture, minus the background, was calculated to generate light curves. The final aperture size for each dataset was selected on the basis of highest resulting signal-to-noise ratio.

Normalized light curves from each of the sites are displayed in Fig. 2. Data from the Clay telescope had the best signal-to-noise ratio (273) and the best time resolution (0.1 s). Although Gemini South is a larger telescope, the throughput was limited by an R filter and data were taken at a significantly slower cadence (0.3 s integration time, with ~0.15 s deadtime). The spatial resolutions of the datasets are 2.13 km (Clay), 4.27 km (du Pont), 10.66 km (Gemini South) and 10.67 km (Cerro Armazones). These values extend from less than two to just over eight times the Fresnel scale ($\sqrt{\lambda D/2} = 1.27$ km, where wavelength λ is 720 nm and D is the distance to Charon).

Models for straight-edge diffraction by a limb, integrated over the exposure interval for each frame, were fitted separately to the

immersion and emersion portions of each light curve. Free parameters of the models included background, full scale, geometric limb occultation time, and effective wavelength. As the dominant averaging effect was integration time, the occulted star was assumed to be a point source, and effects due to transfer time and smearing over a large band of wavelengths were not included. The resulting geocentric limb occultation times at each site were used to calculate Charon’s radius. We have assumed that Charon is spherical and reserve asymmetric shape analysis for future work.

We find that the mean radius of Charon is 606 ± 8 km. The formal radius error from the least-squares fit to a strict circular solution is much smaller (0.04 km) because of our excellent time resolution; however, we feel that this is an underestimate given the sparse sampling of the limb and a suggestion of non-circularity from the Gemini South chord length. Further details concerning our calculation of the radius and error bar are available in Supplementary Information. The radius we find is consistent with previous measurements⁵ and the lower limit established from the previous stellar occultation observation⁴. Assuming that Charon is spherical and has

Table 1 | Upper limits on atmosphere candidate gases

Gas	Molecular weight (a.m.u.)	Refractivity* (10 ^{−4})	Scale height† (km)	Number density upper limit‡ (10 ¹³ cm ^{−3})	Column height upper limit‡ (cm am)
CH ₄	16.04	4.37	90 ± 13	2.0	7.8
H ₂ O	18.02	2.50	79 ± 12	3.4	11.4
Ne	20.18	0.67	71 ± 10	11.9	36.0
N ₂	28.01	2.95	51 ± 8	2.3	5.0
CO	28.01	3.33	51 ± 8	2.0	4.4
Ar	39.95	2.83	36 ± 5	2.0	3.1
CO ₂	44.01	4.47	32 ± 5	1.2	1.7
Kr	83.80	4.25	17 ± 3	0.9	0.7
Xe	131.29	6.97	11 ± 2	0.4	0.2

* At standard temperature and pressure, for wavelengths ranging from 6,438 to 7,056 Å (ref. 11).
† Assuming Charon is spherical, with radius 606 ± 8 km, mass $(1.60 \pm 0.12) \times 10^{24}$ g, and surface temperature 50 K.
‡ Upper limits are 3σ , for an isothermal atmosphere. The number density limit applies to surface density. The upper limits on column height are written in units of centimetre amagats (cm am).

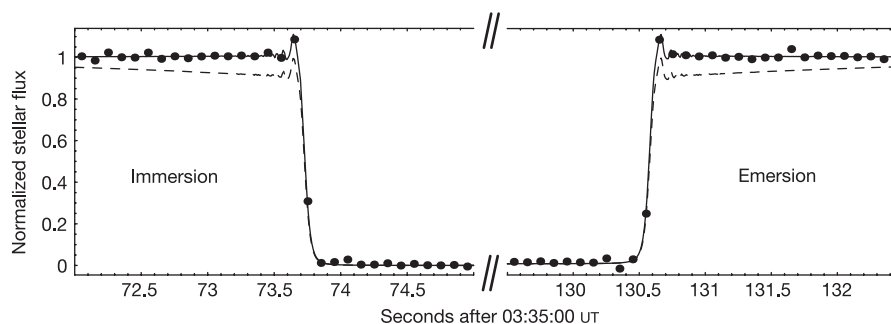


Figure 3 | Atmospheric model fits to the light-curve data. The segments of the Clay light curve shown here have been expanded such that individual data points are resolved, and the first diffraction fringe is clearly visible during immersion and emersion. A diffraction model with a tenuous atmosphere (solid line), from which our 3σ upper limit is derived, is displayed. To illustrate the effect of a substantial atmosphere, a second

model is also shown (dashed line), which represents a 50-km scale height and a 10% flux drop at the surface ($b = 0.10$). This model merges with the data baseline beyond the limits of the plot. The differential bending in the dashed-line model has been enhanced significantly with respect to our upper limit ($b = 0.015$).

a mass of $(1.60 \pm 0.12) \times 10^{24}$ g (ref. 2), we find its density to be 1.72 ± 0.15 g cm $^{-3}$. For ice density 1.0 g cm $^{-3}$ and rock density 3.0 g cm $^{-3}$, we derive a rock-mass fraction of 0.63 ± 0.05 . This value suggests that Charon contains a smaller fraction of rock by mass than Pluto (0.73 ± 0.06 ; refs 2, 5) and Neptune's moon Triton (0.77 ± 0.01 ; ref. 8).

We utilized the detection of the first diffraction fringe in the Clay data to perform model fitting for a thin, isothermal atmosphere in the presence of limb diffraction^{9,10}. Additional parameters for this model included atmospheric scale height and differential bending (H and b)⁹, in the regime where H is much greater than the Fresnel scale. For small flux drops, the differential bending parameter b equals the drop in normalized flux at the surface radius. We selected several likely atmospheric constituents⁴, listed in Table 1. The diffraction models were run over the range of scale heights appropriate for these candidates. Based on the lack of significant flux drop in the data (indicative of no significant atmosphere), we set a 3σ upper limit for differential bending of $b = 0.015$. Figure 3 shows the tenuous-atmosphere model from which this limit was derived (solid line), along with data from the Clay light curve to which it was fitted. The dashed-line model in Fig. 3 illustrates the effect of an atmosphere an order of magnitude denser than the upper limit of the fitted model. From our upper limit on b , we derive a 3σ upper limit on the refractivity for an atmosphere composed of each candidate gas⁴. Refractivities of the gases at standard temperature and pressure¹¹ were then used to calculate upper limits on number density and column height.

The 3σ upper limits on number density for each of the prospective atmospheric constituents are listed in Table 1. The corresponding column heights are roughly a factor of two lower than those established from the previously observed stellar occultation by Charon⁴. Such low densities are consistent with Charon's anticipated high atmospheric loss rate and lack of replenishment mechanism^{12,13}. In context, the upper limits found here are orders of magnitude less than the density of N_2 near the surface of Triton ($\sim 10^{15}$ cm $^{-3}$; ref. 14) and at a radial distance of 1,205 km (the 3- μ bar pressure level in 2002) on Pluto (2.2×10^{14} cm $^{-3}$; refs 14–16). More apt comparisons are the tenuous atmospheres of Io, which is primarily SO_2 and has a number density of a few times 10^{10} cm $^{-3}$ (ref. 17), and Mercury, with a total subsolar atmospheric density of $< 10^7$ cm $^{-3}$ (ref. 18). Leading candidate gases of which a tenuous atmosphere on Charon may be composed are H_2O , N_2 , Ar (ref. 19) and CO (ref. 20). Spectroscopic observations of Charon's surface indicate the presence of H_2O ice^{21,22} mixed with a neutral absorber²³. Other studies have suggested that volatiles could be present on the surface²⁴, perhaps located in shadowed regions¹³. However, the low vapour pressure of H_2O renders it an unlikely atmospheric constituent⁴ and recent spectral

analyses find no indication of volatile ices such as CO, CH_4 or N_2 on Charon's surface²³. The absence of an obvious atmospheric candidate gas is consistent with the lack of a significant atmosphere in the occultation data.

The rock-mass fraction we find for Charon, in addition to those of Pluto and Triton, is higher than the maximum (~ 0.5) predicted by models of outer solar nebula condensates⁸. One explanation is a collisional origin for Pluto–Charon, during which there was a possibility for jetting of icy mantle material⁶. Charon's effective lack of atmosphere could also be explained by a volatile-removing collision⁶. The density we find for Charon is lower than the system mean, which seems consistent with an impact if either precursor object were differentiated⁶. In this case, an impactor could have collided with a differentiated proto-Pluto, leaving a disk of preferentially lower-density material that resulted in a satellite depleted in heavier elements (a situation that has a parallel in the Earth–Moon system)^{6,25}.

This successful observation of a stellar occultation by Charon is encouraging for establishing sizes and probing for atmospheres of large Kuiper belt objects, such as the newly discovered 2003 UB₃₁₃, 2005 FY₉, and 2003 EL₆₁ (ref. 26). In particular, 2003 UB₃₁₃ has approximately Charon's angular diameter and is a prime candidate for having a bound atmosphere, because methane has been detected on its surface²⁷ and it is larger than Pluto.

Received 2 August; accepted 21 September 2005.

1. Null, G. W. & Owen, W. M. J. Charon/Pluto mass ratio obtained with HST CCD observations in 1991 and 1993. *Astron. J.* **111**, 1368–1381 (1996).
2. Olkin, C. B., Wasserman, L. H. & Franz, O. G. The mass ratio of Charon to Pluto from Hubble Space Telescope astrometry with Fine Guidance Sensors. *Icarus* **164**, 254–259 (2003).
3. Walker, A. R. An occultation by Charon. *Mon. Not. R. Astron. Soc.* **192**, 47p–50p (1980).
4. Elliot, J. L. & Young, L. A. Limits on the radius and a possible atmosphere of Charon from its 1980 stellar occultation. *Icarus* **89**, 244–254 (1991).
5. Tholen, D. J. & Buie, M. W. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 193–219 (Univ. Arizona Press, Tucson, 1997).
6. McKinnon, W. B. On the origin of the Pluto–Charon binary. *Astrophys. J.* **344**, L41–L44 (1989).
7. McDonald, S. W. & Elliot, J. L. Pluto–Charon stellar occultation candidates: 2000–2009. *Astron. J.* **119**, 1999–2007 (2000); Erratum. **120**, 1599 (2000).
8. McKinnon, W. B., Simonelli, D. P. & Schubert, G. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 295–343 (Univ. Arizona Press, Tucson, 1997).
9. French, R. G. & Gierasch, P. J. Diffraction calculation of occultation light curves in the presence of an isothermal atmosphere. *Astron. J.* **81**, 445–451 (1976).
10. Bartholdi, P. & Owen, F. The occultation of Beta Scorpii by Jupiter and Io. II. Io. *Astron. J.* **77**, 60–65 (1972).
11. National Research Council (U.S.). *International Critical Tables of Numerical Data, Physics, Chemistry, And Technology* (eds West, C. J. & Hull, C.) (McGraw-Hill Book Co., New York, 1933).

12. Trafton, L., Stern, S. A. & Gladstone, G. R. The Pluto-Charon system: The escape of Charon's primordial atmosphere. *Icarus* **74**, 108–120 (1988).
13. Yelle, R. V. & Elliot, J. L. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 347–390 (Univ. Arizona Press, Tucson, 1997).
14. Elliot, J. L., Person, M. J. & Qu, S. Analysis of stellar occultation data. II. Inversion, with application to Pluto and Triton. *Astron. J.* **126**, 1041–1079 (2003).
15. Elliot, J. L. *et al.* The recent expansion of Pluto's atmosphere. *Nature* **424**, 165–168 (2003).
16. Pasachoff, J. M. *et al.* The structure of Pluto's atmosphere from the 2002 August 21 stellar occultation. *Astron. J.* **129**, 1718–1723 (2005).
17. Spencer, J. *et al.* Mid-infrared detection of large longitudinal asymmetries in Io's SO₂ atmosphere. *Icarus* **176**, 283–304 (2005).
18. Broadfoot, A. L., Shemansky, D. E. & Kumar, S. Mariner 10: Mercury atmosphere. *Geophys. Res. Lett.* **3**, 577–580 (1976).
19. Stern, S. A. & Trafton, L. Constraints on bulk composition, seasonal variation, and global dynamics of Pluto's atmosphere. *Icarus* **57**, 231–240 (1984).
20. Yelle, R. V. & Lunine, J. I. Evidence for a molecule heavier than methane in the atmosphere of Pluto. *Nature* **339**, 288–290 (1989).
21. Marcialis, R. L., Rieke, G. H. & Lebofsky, L. A. The surface composition of Charon: Tentative identification of water ice. *Science* **237**, 1349–1351 (1987).
22. Buie, M. W., Cruikshank, D. P., Lebofsky, L. A. & Tedesco, E. F. Water frost on Charon. *Nature* **329**, 522–523 (1987).
23. Buie, M. & Grundy, W. M. The distribution and physical state of H₂O on Charon. *Icarus* **148**, 324–339 (2000).
24. Roush, T. L. Charon: more than water ice? *Icarus* **108**, 243–254 (1994).
25. Canup, R. M. A giant impact origin of Pluto-Charon. *Science* **307**, 546–550 (2005).
26. Brown, M. E., Trujillo, C. A. & Rabinowitz, D. L. 2003 EL₆₁, 2003 UB₃₁₃, and 2005 FY₉. *IAU Circ.* **8577** (2005).
27. Trujillo, C. A., Barkume, K. M., Brown, M. E., Schaller, E. L. & Rabinowitz, D. L. Near infrared spectra from Mauna Kea of the new brightest Kuiper Belt Object. *Bull. Am. Astron. Soc.* **37**, 52.06 (2005).
28. USNO. *The Astronomical Almanac for the Year 2005* (US Govt Printing Office, Washington DC, 2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements B.A.B., J.W.G. and T.T. are Guest Observers at Instituto de Astronomía, Universidad Católica del Norte, Antofagasta, Chile. We thank N. Vogt and M. Murphy of Universidad Católica del Norte for their expertise, assistance, and the use of their telescope at Cerro Armazones; G. Gutiérrez and F. Sánchez, telescope operators at Las Campanas; R. Carrasco, K. Volk (observers), and E. Wendroth (telescope operator) at Gemini South; M. Ottoboni and R. Campos at Pico Dos Dias; and C. Czelusniak, assistant observer from Universidade Estadual de Ponta Grossa. Full credits for Gemini Observatory are available at <http://www.us-gemini.noao.edu/sciops/data/dataAcknowIndex.html>. Support for this work was provided by NASA Planetary Astronomy. Additional support was provided by the Belgian Federal Office for Scientific, Technical and Cultural Affairs and the Flemish Ministry for Foreign Policy, European Affairs, Science and Technology.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.A.S.G. (gulbis@mit.edu).

LETTERS

Charon's size and an upper limit on its atmosphere from a stellar occultation

B. Sicardy^{1,2}, A. Bellucci¹, E. Gendron¹, F. Lacombe¹, S. Lacour¹, J. Lecacheux¹, E. Lellouch¹, S. Renner¹, S. Pau¹, F. Roques¹, T. Widemann¹, F. Colas³, F. Vachier³, R. Vieira Martins^{3,15}, N. Ageorges⁴, O. Hainaut⁴, O. Marco⁴, W. Beisker⁵, E. Hummel⁵, C. Feinstein⁶, H. Levato⁷, A. Maury⁸, E. Frappa⁹, B. Gaillard¹⁰, M. Lavyassière¹⁰, M. Di Sora¹¹, F. Mallia¹¹, G. Masi^{11,12}, R. Behrend¹³, F. Carrier¹³, O. Mousis¹⁴, P. Rousselot¹⁴, A. Alvarez-Candal¹⁵, D. Lazzaro¹⁵, C. Veiga¹⁵, A. H. Andrei^{15,16}, M. Assafin¹⁶, D. N. da Silva Neto¹⁶, C. Jacques¹⁷, E. Pimentel¹⁷, D. Weaver¹⁸, J.-F. Lecampion¹⁹, F. Doncel²⁰, T. Momiyama²⁰ & G. Tancredi²¹

Pluto and its satellite, Charon (discovered in 1978; ref. 1), appear to form a double planet, rather than a hierarchical planet/satellite couple. Charon is about half Pluto's size and about one-eighth its mass. The precise radii of Pluto and Charon have remained uncertain, leading to large uncertainties on their densities². Although stellar occultations by Charon are in principle a powerful way of measuring its size, they are rare, as the satellite subtends less than 0.3 microradians (0.06 arcsec) on the sky. One occultation (in 1980) yielded a lower limit of 600 km for the satellite's radius³, which was later refined to 601.5 km (ref. 4). Here we report observations from a multi-station stellar occultation by Charon, which we use to derive a radius, $R_C = 603.6 \pm 1.4$ km (1σ), and a density of $\rho = 1.71 \pm 0.08$ g cm⁻³. This occultation also provides upper limits of 110 and 15 (3σ) nanobar for an atmosphere around Charon, assuming respectively a pure nitrogen or pure methane atmosphere.

Charon occulted the 15th magnitude star UCAC2 26257135 on 11 July 2005, as initially predicted by D. Herald (personal communication). Charon's occultation shadow swept South America, where some of the largest telescopes in the world were available. Table 1 provides the timing of the occultation at three stations, yielding kilometre-level accuracy on the length of the occultation segments (or 'chords') at each station, using Charon's shadow velocity. We performed circular fits to the chord extremities (the three red segments in Fig. 1), the three free parameters being the two coordinates of Charon's centre and its radius. The chord extremities were weighted according to the uncertainties in the occultation times, converted into radial uncertainties, perpendicular to Charon's limb. The best fit yields a standard radial deviation of 1.1 km, and a χ^2 value per degree of freedom of 0.85 (Table 2), indicating a satisfactory fit. The corresponding radius of Charon is $R_C = 603.6 \pm 1.4$ km (formal 1σ error), assuming that the limb is circular, that is, that there are only three free parameters to adjust.

Our data do not reveal significant departures from circularity. Although elliptical fits do improve the residuals, they also reduce the number of degrees of freedom of the fit, by adding the oblateness and

ellipse orientation as free parameters. This eventually worsens the χ^2 per degree of freedom (Table 2), but also increases the formal error bar on R_C from 1.4 to 5 km. We obtain a 1σ upper limit of 8×10^{-3} for the limb oblateness, fifty times larger than the value expected for a slow 6.4-day rotator in hydrostatic equilibrium. Furthermore, local topographic features might alter our determination of R_C by a few kilometres. Larger features (height >10 km) are not expected to occur, as they should relax over geologic timescales owing to the structural weakness of methane and nitrogen ices⁵. Also, our measurements apply to Charon's shape projected in the instantaneous plane of the sky, with no access to other planes. All considered, however, the global uncertainty on R_C should be smaller than 5 km. Finally, in the presence of a tenuous atmosphere, the stellar rays would be refracted towards the Earth, resulting in a shadow slightly reduced compared to Charon's body (see below).

Our result comes after two decades of extensive discussions on Charon's radius⁶. The values derived from the mutual events—occultations and eclipses of Pluto by Charon and vice versa—observed in the 1980s range from $R_C = 590 \pm 5$ km, to 592 ± 13 km, 611 ± 30 km and 627 ± 21 km (refs 7, 8, 9 and 10, respectively, 1σ error bars), assuming a semimajor axis of 19,599 km for Charon. They are thus all within 1.2σ of our value (except for the first value, at 2.7σ). Their differences mainly reflect the use of different data sets, and in some cases, of different modelling (albedo features or limb darkening).

A recent, improved orbit for Charon includes observations with the Hubble Space Telescope^{2,11,12}, besides older measurements made since 1978. The physical parameters used for this orbit are, among others (R. A. Jacobson, personal communication): total mass of the system $M = (1.463 \pm 0.0033) \times 10^{22}$ kg, mass ratio Charon/Pluto $f = 0.121 \pm 0.006$, semimajor axis $a = 19,599.0 \pm 15$ km. This provides Charon's mass $m_C = (1.58 \pm 0.07) \times 10^{21}$ kg, where most of the error bar comes from the uncertainty on f . Combining this mass with our value of R_C yields Charon's density $\rho_C = 1.71 \pm 0.08$ g cm⁻³, where most of the error bar comes from the uncertainty on Charon's mass, as its volume is now accurately

¹Observatoire de Paris, LESIA, 92195 Meudon cedex, France. ²Université Pierre et Marie Curie, 75252 Paris cedex 5, France. ³Observatoire de Paris, IMCCE, 75014 Paris, France.

⁴European Southern Observatory, Alonso de Córdova 3107, Casilla 19001, Santiago 19, Chile. ⁵International Occultation Timing Association, European Section, 30459 Hannover, Germany. ⁶Facultad de Ciencias Astronómicas y Geofísicas, Observatorio Astronómico & Instituto de Astrofísica de La Plata, CONICET, Paseo del Bosque 1900 La Plata, Argentina. ⁷Complejo Astronómico, El Leoncito, CP J5402DSP, San Juan, Argentina. ⁸Gene Shoemaker Observatory, Casilla 21, San Pedro de Atacama, Chile. ⁹Planétarium de Saint-Etienne, 42100 Saint-Etienne, France. ¹⁰Association des Utilisateurs de Détecteurs Electroniques (AUDE), France, c/o F. Colas, 45, Av. Reille, 75014 Paris, France. ¹¹Campo Catino Austral Observatory, Casilla 21, San Pedro de Atacama, Chile. ¹²Università di Tor Vergata di Roma, Via della Ricerca Scientifica n.1, 00133, Rome, Italy. ¹³Observatoire de Genève, CH-1290 Sauverny, Switzerland. ¹⁴Observatoire de Besançon, BP1615, 25010 Besançon cedex, France. ¹⁵Observatório Nacional, 20921-400, Rio de Janeiro, Brazil.

¹⁶Observatório do Valongo/UFRJ, CEP 20080-090, Rio de Janeiro, Brazil. ¹⁷Observatório CEAMIG-REA, CEP 31545-120, Belo Horizonte, MG, Brazil. ¹⁸Observatório Astronômico Christus, Universidade de Fortaleza, rua João Carvalho, 630, CEP 60140-140 Fortaleza, Brazil. ¹⁹Observatoire Aquitain des Sciences de l'Univers, 33270 Floirac, France. ²⁰Observatorio Astronómico, Universidad Nacional de Asunción 2169, Paraguay. ²¹Observatorio Astronómico Los Molinos, Facultad de Ciencias, 11400 Montevideo, Uruguay.

Table 1 | Circumstances of observations for the 11 July 2005 Charon occultation

Site*	Telescope, cycle time, effective wavelength	Latitude, longitude, altitude	Disappearance†, re-appearance† (h:min:s, UT, 11 July 2005)	Shadow velocity (km s ⁻¹)
San Pedro de Atacama	'Campo Catino Austral Telescope' (0.5 m), 0.716 s, 0.65 μ m	68° 10' 48.2" W, 22° 57' 08.4" S, 2,410 m	03:36:20.98 $\pm$ 0.18, 03:36:28.30 $\pm$ 0.30	21.347, 21.347
Paranal	'Yepun' VLT (8.2 m), 0.2 s, 2.2 μ m	70° 24' 07.9" W, 24° 37' 31.0" S, 2,635 m	03:36:18.09 $\pm$ 0.04, 03:36:55.40 $\pm$ 0.05	21.345, 21.345
El Leoncito	'Jorge Sahade Telescope' (2.15 m), 1 s, 0.7 μ m	69° 17' 44.9" W, 31° 47' 55.6" S, 2,492 m	03:36:15.03 $\pm$ 0.16, 03:37:02.98 $\pm$ 0.08	21.317, 21.317

*We attempted observations of the Charon occultation from Argentina, Bolivia, Brazil, Chile, Paraguay and Uruguay. Owing to weather conditions or technical problems, not all the stations recorded the event. The present paper is based on data gathered at San Pedro de Atacama (Chile), at Cerro Paranal (Chile) with the Very Large Telescope (VLT) of the European Southern Observatory, and at El Leoncito (Argentina), listed here. Whereas observations at both San Pedro and El Leoncito were made with fast broadband visible CCD, the Paranal observations were achieved with the NACO adaptive optics camera using a K_s band filter (2.2 μ m). In the latter case, we were able to resolve the Pluto/Charon pair, with the two objects separated by 0.89 arcsec during the occultation. Beyond the three stations listed above, the occultation was also observed from La Silla (Chile) with the 1.2-m swiss telescope in drift scan mode, but at irregular speed, making the use of the light curve impossible in this paper. We furthermore obtained data from Asunción (Paraguay) with a 0.45-m telescope and a broadband CCD detector. Owing to their large cycle time (7 s), however, these data are not included in this analysis. Images were finally acquired at the CEAMIG-REA 0.3-m telescope in Belo Horizonte (Brazil), under partly cloudy conditions and with poor signal-to-noise ratio, making this data set unusable for the present analysis.

†The disappearance and reappearance times are obtained by fitting an abrupt edge shadow to the light curves, after convolving the shadow by Fresnel diffraction, stellar diameter (0.42 km projected at Charon) and finite integration time of the instrument. The error bars on the timings are 1 σ level (68.3% confidence level) provided by those fits.

determined. This is true as long as the uncertainty on R_C remains smaller than 10 km, a safe margin, as discussed earlier.

Comparison with Pluto's density is problematical, however, as the planet radius is not so accurately determined. Owing to refraction by Pluto's atmosphere, occultation determination of Pluto's radius, R_P , still depends on atmospheric models¹³. An upper limit of $R_P = 1,195 \pm 5$ km is given by occultations¹⁴, while a lower limit of $R_P = 1,151 \pm 6$ km is provided by mutual events⁶. Combining these results with Pluto's mass, derived from the quantities above, yields Pluto's density in the range 1.8–2.1 g cm⁻³ (ref. 2), where most

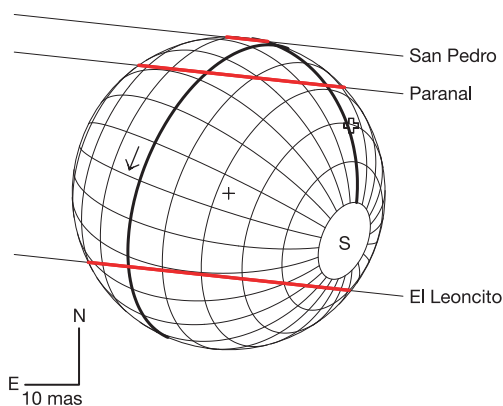


Figure 1 | Measuring Charon's radius. Charon's aspect on 11 July 2005, with celestial north up and east left, using the values in Table 2. The scale in milli-arcsec (mas) is shown, with one mas corresponding to 21.809 km projected at Charon. The thicker meridian is the origin of longitudes on Charon, that is, the meridian always facing Pluto, as the satellite is locked in a synchronous orbit. The thicker parallel is the equator. Charon's south pole (S) follows the IAU definition, the arrow indicating the satellite rotation. The star trajectories relative to Charon, as observed from San Pedro, Paranal and El Leoncito, are shown as black lines, the red parts corresponding to the segments where the star was occulted by Charon. A circular fit to these chords yields a radius of 603.6 ± 1.4 km (1 σ) for Charon. The thick cross marks the expected location of Charon's centre, using the DE413/PLU013 Charon ephemeris, and the ICRF/J2000 star position given in Table 2. The thin cross is the centre of the circular fit, showing that Charon's DE413/PLU01 position must be corrected by $\Delta\alpha\cos(\delta) = +22 \pm 11$ mas (towards the east) and $\Delta(\delta) = -12 \pm 11$ mas (towards the south), where the error bars come from the uncertainties on the star position. This offset is mostly attributable to an error on Pluto's barycentric DE413 ephemeris, rather than to an offset of Charon's PLU013 ephemeris around Pluto. In fact, adaptive optics images taken with the NACO/VLT camera at Paranal show that Charon is at only 4 mas from its calculated position relative to Pluto, an effect that could be entirely due to photocentre displacements caused by albedo features on Pluto and/or Charon.

of the uncertainty now comes from Pluto's radius R_P , not from its mass. However, our results tighten the difference between Pluto's density and Charon's, as the latter was previously estimated² to lie in the interval 1.4–1.8 g cm⁻³.

These ranges for Pluto's and Charon's densities are in good agreement with current structural models¹⁵, which produce baseline densities of 1.85 g cm⁻³ and 1.75 g cm⁻³ for Pluto and Charon, respectively. They indicate a slightly higher rock versus ice fraction on Pluto (0.65) than on Charon (0.55–0.60). Our improved density for Charon, however, cannot distinguish differentiated and undifferentiated states of the satellite. In the framework of the giant impact model for the origin of Pluto and Charon, similar densities for the two bodies favour the scenario in which Charon is formed intact, as opposed to being accreted from a disk orbiting Pluto¹⁶.

Note that there is now a possibility of improving these numbers by re-analysing the mutual events of the 1980s, using the value of Charon's radius derived here, plus the improved orbital parameters quoted above, in order to get a more accurate value for R_P . This

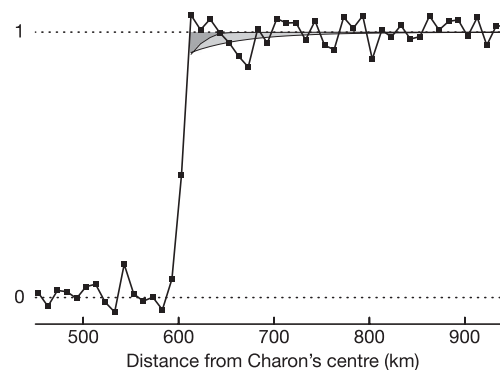


Figure 2 | Limit on Charon's atmosphere. The stellar flux from Leoncito and Paranal before and after the occultation has been rebinned in intervals of 10 km in radial distance from Charon's centre. The two data sets have then been averaged with weights taking into account their respective noise levels, resulting in the light curve shown here (black squares connected by a line). The values have been normalized between zero (no stellar flux) and unity (full stellar flux), as indicated by the dotted lines. Two examples of atmospheric models are shown superimposed on the data. Light grey model: expected drop of signal with an isothermal N₂ atmosphere at $T = 56$ K, with a pressure of $p_s = 110$ nbar at Charon's surface. Dark grey model: effect of a CH₄ atmosphere with $T = 56$ K and $p_s = 15$ nbar at the surface, with T increasing to 100 K near 20 km above the surface, thus mimicking Pluto's atmosphere temperature profile. These models illustrate upper limits of detection (at 3 σ level) that we can obtain on a putative atmosphere for Charon.

Table 2 | Fits to the occultation chords

Site	f^* (km)	g^* (km)	Latitude† of suboccultation point (deg.)	Radial residual (km)		
				Circular fit, $\epsilon = 0$ fixed	Elliptical fit, P fixed	Elliptical fit, P free
San Pedro, disappearance	+327.1	+325.0	06.9 N	+0.26	+0.40	+0.31
San Pedro, re-appearance	+482.6	+341.6	19.1 N	−0.84	−0.75	−0.92
Paranal, disappearance	+28.7	+147.6	20.5 S	−0.14	−0.22	−0.08
Paranal, re-appearance	+820.6	+232.5	44.2 N	+0.28	+0.07	+0.11
El Leoncito, disappearance	−2.8	−636.7	52.7 S	+2.45	+1.93	+0.26
El Leoncito, re-appearance	+1,013.5	−527.6	33.2 N	−0.60	−0.25	−0.07
Free parameters				Best-fit values		
Charon's radius, R_C (km)				603.6	603.1	603.4
Offset‡ in right ascension, f_c (km)				472.7	472.4	+471.9
Offset‡ in declination, g_c (km)				−261.0	−260.8	+261.8
Oblateness, ϵ				0, fixed	-1.5×10^{-3}	-2×10^{-3}
North pole position angle, P (deg.)				67.6, fixed	67.6, fixed	+33.3
χ^2 per degree of freedom§				0.85	1.10	1.67

*The timings of Table 1 provide the star position relative to Charon's expected centre, using the DE413/PLU013 Charon ephemeris (<http://ssd.jpl.nasa.gov>). This position is projected in the plane of the sky, in km, where f is the relative position in right ascension, positive if the star is east of Charon's centre, and g is the relative position in declination, positive if the star is north of Charon's centre. We used the following ICRF/J2000 star position: $\alpha = 17^{\text{h}} 28^{\text{m}} 55.0167^{\text{s}}$ and $\delta = -15^{\circ} 00' 54.726''$, with typical uncertainties of 11 mas, measured at the 60-cm reflector of Pico dos Dias (Laboratório Nacional de Astrofísica, Brazil), and at the meridian refractor of Bordeaux Observatory (France).

†The latitudes of the suboccultation points on Charon are derived using a north pole position angle of $P = 67.6^{\circ}$ with respect to the J2000 celestial north, and a sub-Earth latitude of $B = -34.2^{\circ}$.

‡This offset is the position of Charon's centre obtained from the fit (thin cross in Fig. 1), relative to Charon's centre expected from the adopted star position and the DE413/PLU013 ephemeris (thick cross in Fig. 1).

§The number of degrees of freedom is the number of data points (here $N = 6$) minus the number of free parameters: $M = 3$, $M = 4$ or $M = 5$, depending on whether the fit is circular, elliptical with P fixed, or elliptical with P free, respectively. The quantity minimized in the fits is $\chi^2 = \sum_i (r_{i,\text{obs}} - r_{i,\text{cal}})^2 / \sigma_i^2$, where $r_{i,\text{obs}}$ (resp. $r_{i,\text{cal}}$) is the distance of the observed (resp. calculated) i th point to the shadow centre, and σ_i is the 1σ uncertainty on $r_{i,\text{obs}}$.

would have important consequences for better constraining not only Pluto's density, but also the Pluto atmosphere models, through a reassessment of occultation observations.

Our data also set an upper limit for a putative atmosphere for Charon. By combining the stellar fluxes observed at the Paranal and El Leoncito observatories, we derive a synthetic light curve, as shown in Fig. 2. The effect of an atmosphere depends on the surface pressure, the nature of the gas and the temperature profile. We assumed two cases. One is that of an isothermal nitrogen (N_2) atmosphere at $T_s = 56$ K, the recently estimated mean daytime Charon surface temperature¹⁷. The other is a pure methane (CH_4) atmosphere, with a temperature increasing from 56 K at the surface to 100 K above 20 km, due to solar heating, as is the case for Pluto's atmosphere¹⁴. The two cases indicate upper limits of 110 and 15 nbar (3σ), respectively, with corresponding upper limits of 4.1 and 1.3 cm amagat for the vertical column densities. Limits obtained from the 1980 Charon stellar occultation were about two and ten times larger for N_2 and CH_4 , respectively⁴. Note that in the limiting cases presented here, refraction of stellar rays by the atmosphere would cause a reduction of Charon's shadow radius by about 10 km, when compared to the actual radius, R_C . Consequently, if an atmosphere is detected at those levels in the future, such effects should be considered when deriving R_C .

The very low upper limit for an atmosphere around Charon is not surprising, given estimates of escape rates¹⁴. The upper limit we derive for a pure methane atmosphere is also consistent with the absence of a CH_4 ice signature in its near-infrared spectrum¹⁸. In fact, a 15 nbar CH_4 atmosphere is in equilibrium with CH_4 ice at 41 K, much less than the 56 K quoted above. Methane ice could still be present in restricted, colder, regions of the surface. For N_2 , a 110 nbar atmosphere would imply an even lower equilibrium temperature ($T < 31$ K), requiring that N_2 ice be confined at best to high northern latitudes and/or to permanently shadowed regions of the satellite. The same is true for other candidates, like CO, which would require temperatures as low as 35 K.

Received 2 September; accepted 17 October 2005.

- Christy, J. W. & Harrington, R. S. The satellite of Pluto. *Astron. J.* **83**, 1005–1008 (1978).

- Olkin, C. B., Wasserman, L. H. & Franz, O. G. The mass ratio of Charon to Pluto from Hubble Space telescope astrometry with the fine guidance sensors. *Icarus* **164**, 254–259 (2003).
- Walker, A. R. An occultation by Charon. *Mon. Not. R. Astron. Soc.* **192**, 47P–50P (1980).
- Elliot, J. L. & Young, L. A. Limits on the radius and a possible atmosphere of Charon from its 1980 stellar occultation. *Icarus* **89**, 244–254 (1991).
- Stern, S. A. The Pluto-Charon system. *Annu. Rev. Astron. Astrophys.* **30**, 185–233 (1992).
- Tholen, D. J. & Buie, M. W. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 193–219 (Univ. Arizona Press, Tucson, 1997).
- Reinsch, K., Burwitz, V. & Festou, M. C. Albedo maps of Pluto and improved physical parameters of the Pluto-Charon system. *Icarus* **108**, 209–218 (1994).
- Tholen, D. J. & Buie, M. W. Further analysis of the Pluto-Charon mutual event observations. *Bull. Am. Astron. Soc.* **22**, 1129 (1990).
- Buratti, B. J. et al. Modeling Pluto-Charon mutual events. II. CCD observations with the 60 in. telescope at Palomar Mountain. *Astron. J.* **110**, 1405–1419 (1995).
- Young, E. F. & Binzel, R. P. A new determination of radii and limb parameters for Pluto and Charon from mutual events lightcurves. *Icarus* **108**, 219–224 (1994).
- Null, G. W. & Owen, W. M. Jr Charon/Pluto mass ratio obtained with HST CCD observations in 1991 and 1993. *Astron. J.* **111**, 1368–1381 (1996).
- Tholen, D. J. & Buie, M. W. The orbit of Charon. I. New Hubble Space Telescope observations. *Icarus* **125**, 245–260 (1997).
- Stansberry, J. A., Lunine, J. I., Hubbard, W. B., Yelle, R. V. & Hunten, D. M. Mirages and the nature of Pluto's atmosphere. *Icarus* **111**, 503–513 (1994).
- Yelle, R. V. & Elliot, J. L. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 347–390 (Univ. Arizona Press, Tucson, 1997).
- McKinnon, W. B., Simonelli, S. P. & Schubert, G. in *Pluto and Charon* (eds Stern, S. A. & Tholen, D. J.) 295–343 (Univ. Arizona Press, Tucson, 1997).
- Canup, R. M. A giant impact origin of Pluto-Charon. *Science* **307**, 546–550 (2005).
- Gurwell, M. A. & Butler, B. J. Sub-arcsec scale imaging of the Pluto/Charon binary system at 1.4 mm. *Bull. Am. Astron. Soc.* **37**, 743 (2005).
- Dumas, C., Terrile, R. J., Brown, R. H., Schneider, G. & Smith, B. A. Hubble Space Telescope NICMOS spectroscopy of Charon's leading and trailing hemispheres. *Astron. J.* **121**, 1163–1170 (2001).

Acknowledgements We thank the Conseil Scientifique of the Paris Observatory and the Programme National de Planétologie for supporting part of the observations of this event in South America.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.S. (bruno.sicardy@obspm.fr).

Structural diversity in binary nanoparticle superlattices

Elena V. Shevchenko^{1,2,*†}, Dmitri V. Talapin^{1,*†}, Nicholas A. Kotov³, Stephen O'Brien² & Christopher B. Murray¹

Assembly of small building blocks such as atoms, molecules and nanoparticles into macroscopic structures—that is, ‘bottom up’ assembly—is a theme that runs through chemistry, biology and material science. Bacteria¹, macromolecules² and nanoparticles³ can self-assemble, generating ordered structures with a precision that challenges current lithographic techniques. The assembly of nanoparticles of two different materials into a binary nanoparticle superlattice (BNSL)^{3–7} can provide a general and inexpensive path to a large variety of materials (metamaterials) with precisely controlled chemical composition and tight placement of the components. Maximization of the nanoparticle packing density has been proposed as the driving force for BNSL formation^{3,8,9}, and only a few BNSL structures have been predicted to be thermodynamically stable. Recently, colloidal crystals with micrometre-scale lattice spacings have been grown from oppositely charged polymethyl methacrylate spheres^{10,11}. Here we demonstrate formation of more than 15 different BNSL structures, using combinations of semiconducting, metallic and magnetic nanoparticle building blocks. At least ten of these colloidal crystalline structures have not been reported previously. We demonstrate that electrical charges on sterically stabilized nanoparticles determine BNSL stoichiometry; additional contributions from entropic, van der Waals, steric and dipolar forces stabilize the variety of BNSL structures.

Face-centred-cubic (f.c.c.) ordering of monodisperse hard spheres dispersed in a liquid permits larger local free space available for each sphere compared to the unstructured phase, resulting in higher translational entropy of the spheres. When the volume fraction of hard spheres approaches ~55%, this ordering enhances the total entropy of the system and drives the ordering phase transition. Entropy-driven crystallization has been studied in great detail both theoretically¹² and experimentally on monodisperse latex particles, whose behaviour can be approximated by hard spheres^{13,14}. In a mixture containing spheres of two different sizes (radii R_{small} and R_{large}), the packing symmetry depends on the size ratio of the small and large spheres ($\gamma = R_{\text{small}}/R_{\text{large}}$)^{3,8}. Calculations show that assembly of hard spheres into binary superlattices isostructural with NaCl, AlB₂ and NaZn₁₃ can be driven by entropy alone without any specific energetic interactions between the spheres^{9,15}. Indeed, NaZn₁₃- and AlB₂-type assemblies of silica particles were found in natural Brazilian opals¹⁶ and can be grown from latex spheres¹⁷. In a certain γ range, the packing density of these structures either exceeds or is very close to the density of the close-packed f.c.c. lattice (0.7405), while structures with lower packing densities are predicted to be unstable^{8,15}.

Despite these predictions, we observed an amazing variety of BNSLs that self-assemble from colloidal solutions of nearly spherical

nanoparticles of different materials (Fig. 1). Coherently packed domains extend up to 10 μm in lateral dimensions, and can display well defined facets (Supplementary Fig. 1). In many cases, several BNSL structures form simultaneously on the same substrate, under identical experimental conditions. The same nanoparticle mixture can assemble into BNSLs with very different stoichiometry and packing symmetry. For example, 11 different BNSL structures were prepared from the same batches of 6.2 nm PbSe and 3.0 nm Pd nanoparticles (Supplementary Fig. 2). We also observe that, in general, BNSLs tolerate much broader γ ranges than hard spheres: for example, AlB₂-type BNSLs assembled from different combinations of PbSe, PbS, Au, Ag, Pd, Fe₂O₃, CoPt₃ and Bi nanoparticles in a broad γ range (Supplementary Fig. 3). Further, we observe BNSLs that could not be identified as isostructural with specific intermetallic compounds (Supplementary Fig. 2). This observed structural diversity of BNSLs defies traditional expectations, and shows the great potential of modular self-assembly at the nanoscale.

The formation of binary structures with packing density significantly lower than the density of single-phase f.c.c. close packing (0.7405) rules out entropy as the main driving force for nanoparticle ordering. Moreover, van der Waals, steric or dipolar interparticle interactions are not sufficient to explain why these low density BNSLs form, instead of their constituents separating into single-component superlattices. Opposite electrical charges on nanoparticles could impart a specific affinity of one type of particle (for example, dodecanethiol-capped Au, Ag, Pd) for another (typically PbSe, PbS, Fe₂O₃, CoPt₃ and so on, capped with long chain carboxylic acids). If nanoparticles are oppositely charged, the Coulomb potential would stabilize the BNSL while destabilizing the single-component superlattices. The electrical charges might be present on sterically stabilized nanoparticles even in non-polar solvents^{18–20}.

To measure charges on the nanoparticles that form our BNSL, we studied the electrophoretic mobility of PbSe and Au nanocrystals. Laser Doppler velocimetry allows the distribution of electrophoretic mobilities within an ensemble of nanoparticles to be measured. The electrical charge (Z , in units of e) of a spherical particle in a low dielectric solvent in absence of electrolyte can be calculated from the electrophoretic mobility (μ_e) where $\mu_e = Ze/(3\pi\eta a)$, η is the viscosity of the solvent and a is the hydrodynamic diameter of a particle²¹. With $a = 10$ nm, we obtain $\mu_e \approx 0.27 \times 10^{-4} Z \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. These calculated values agree well with the peaks in the experimental mobility distribution for 7.2-nm-diameter PbSe nanocrystals in chloroform (Fig. 2a). Owing to the organic coat (oleic acid), the effective hydrodynamic radius of PbSe nanocrystals extends beyond the crystalline core by 1–2 nm, depending on the density of surface coverage. The peaks in the mobility distribution curve indicate the

¹IBM Research Division, T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York 10598, USA. ²Department of Applied Physics & Applied Mathematics, Columbia University, 200 SW Mudd Building, 500 West 120th Street, New York, New York 10027, USA. ³Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA. [†]Present address: The Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

*These authors contributed equally to this work.

presence of particles with charges $-e$, 0 , e and $2e$ in a colloidal solution of monodisperse PbSe nanocrystals.

We found that the charges on PbSe nanocrystals can be altered by adding surfactant molecules like carboxylic acids and tri-*n*-alkylphosphine oxides. Addition of oleic acid increases the population of positively charged PbSe nanocrystals at the expense of the negatively charged and neutral nanocrystals. Depending on the amount of acid added, the majority of nanocrystals can be adjusted to have either one or two positive charges (Fig. 2b and c). Addition of oleic acid increases the solutions' viscosity, causing the peaks to shift towards lower mobility (compare Fig. 2a–c). The addition of tri-*n*-octylphosphine oxide (TOPO) increases the population of negatively

charged PbSe nanocrystals and reduces the concentration of positively charged nanocrystals (Fig. 2d). Surveys of many samples revealed that the additives reliably shifted the distribution of charge states; however, the initial proportion of particles in each charge state was dependent somewhat on sample processing. Both neutral and negatively charged nanoparticles were detected in chloroform solutions of 4.8 nm dodecanethiol-capped Au nanocrystals (Supplementary Fig. 5). After addition of oleic acid most Au nanoparticles become negatively charged (Fig. 2e), whereas the addition of TOPO neutralizes the Au nanoparticles (Fig. 2f). The charges on PbSe and Au nanoparticles could originate from deviations in nanocrystal stoichiometry and adsorption/desorption of charged capping

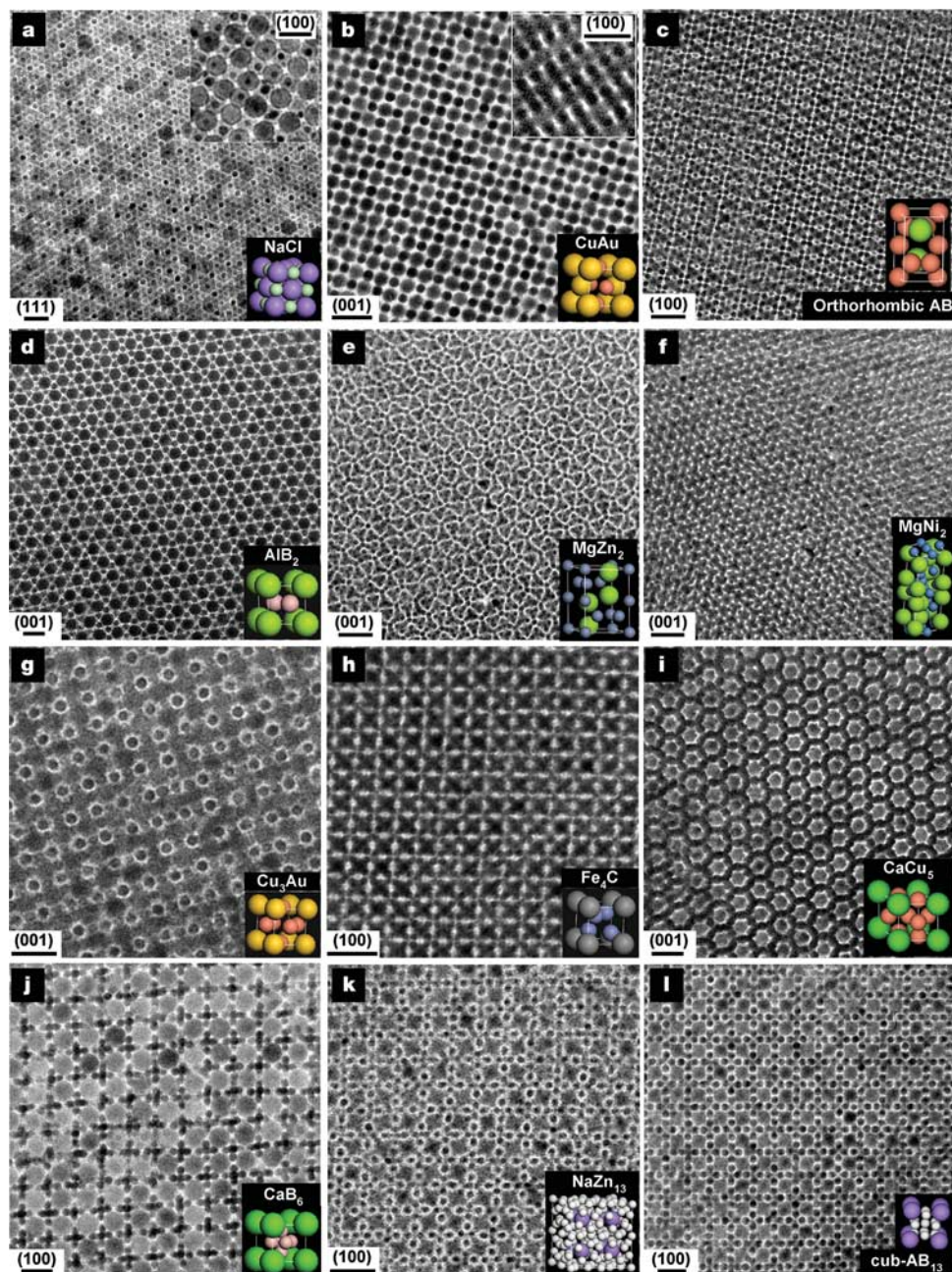


Figure 1 | TEM images of the characteristic projections of the binary superlattices, self-assembled from different nanoparticles, and modelled unit cells of the corresponding three-dimensional structures. The superlattices are assembled from **a**, 13.4 nm γ -Fe₂O₃ and 5.0 nm Au; **b**, 7.6 nm PbSe and 5.0 nm Au; **c**, 6.2 nm PbSe and 3.0 nm Pd; **d**, 6.7 nm PbS and 3.0 nm Pd; **e**, 6.2 nm PbSe and 3.0 nm Pd; **f**, 5.8 nm PbSe and 3.0 nm Pd;

g, 7.2 nm PbSe and 4.2 nm Ag; **h**, 6.2 nm PbSe and 3.0 nm Pd; **i**, 7.2 nm PbSe and 5.0 nm Au; **j**, 5.8 nm PbSe and 3.0 nm Pd; **k**, 7.2 nm PbSe and 4.2 nm Ag; and **l**, 6.2 nm PbSe and 3.0 nm Pd nanoparticles. Scale bars: **a–c**, **e**, **f**, **i–l**, 20 nm; **d**, **g**, **h**, 10 nm. The lattice projection is labelled in each panel above the scale bar. The modelled projections of the binary superlattices are shown in Supplementary Fig. 4.

ligands. Although these additives are effective in adjusting the particle charge states, the specific interactions by which this charge tuning occurs will require further study (Supplementary Discussion 1).

In the presence of oleic acid, PbSe and Au nanoparticles are oppositely charged (Fig. 2e). The Coulomb potential between two oppositely charged nanoparticles ($Z = \pm 1$) separated by 10 nm of a solvent like chloroform is comparable with kT at room temperature, and solutions of mixed PbSe and metal nanoparticles retain stability for several weeks. The relatively small interparticle potential favours annealing of the BNSLs as they grow. For a NaCl-type BNSL with $Z_+ = 1$, $Z_- = -1$ and the nearest-neighbour distance $R_0 = 11.5$ nm (Fig. 1a), the Coulomb binding energy per unit cell is estimated to be $U_{\text{Coul}} \approx MZ_+Z_-e^2/(4\pi\epsilon\epsilon_0R_0) \approx -0.1$ eV (or about $-4kT$ at the superlattice growth temperature, 50°C), where $M = -1.7476$ is the Madelung constant. The Coulomb binding energy is comparable to the van der Waals attractive energy expected for a NaCl-type BNSL. The energy of short-range van der Waals forces ($\sim 1/R^6$) can rival long-range Coulomb energy ($\sim 1/R$) only at the nanometre scale. In BNSLs, we can neglect screening of the Coulomb potential by charged species in solution because the Debye screening length ($\sim 10^{-4}$ cm) is much larger than R_0 (refs 10, 11). In an AB_x BNSL where A and B hold opposite charges, the Coulomb potential per AB_x 'molecule' is $U_{\text{Coul}} \approx -\alpha + \beta(xZ_- + Z_+)^2N^{2/3}$, where α and β are positive constants and N is the number of assembled nanoparticles (Supplementary Discussion 2).

Coulomb energy determines the stoichiometry of the growing BNSL. An extended three-dimensional BNSL can form only if the positive and negative charges compensate each other. If during growth the BNSL accumulates non-compensated charge, eventually U_{Coul} changes sign from negative to positive and the growth is self-limiting. The superlattice nucleation stage should be less sensitive to the Coulomb interactions. Indeed, we observed that many small domains with different BNSL structures can simultaneously nucleate on the same substrate, but their size does not exceed $\sim 10^2$ nanoparticles. Only one or two structures grow to larger length scales ($\sim 10^6$ – 10^8 particles). BNSLs with many particles per unit cell (for

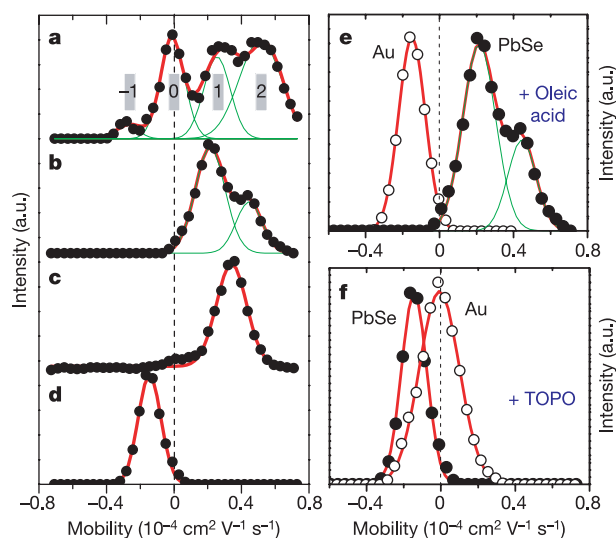


Figure 2 | Electrophoretic mobility of PbSe and Au nanocrystals in chloroform. **a–d**, Distribution of electrophoretic mobility for 7.2 nm PbSe nanocrystals. **a**, PbSe nanocrystals washed to remove excess of capping ligands. The grey bars show mobilities predicted for nanocrystals with charges of -1 , 0 , 1 and 2 (in units of e). **b–d**, Electrophoretic mobility of PbSe nanocrystals in the presence of **b**, 0.02 M oleic acid, **c**, 0.06 M oleic acid and **d**, 0.05 M tri-*n*-octylphosphine oxide. **e, f**, Comparison of electrophoretic mobilities of 7.2 nm PbSe and 4.8 nm Au nanocrystals in the presence of **e**, 0.02 M oleic acid and **f**, 0.05 M tri-*n*-octylphosphine oxide, respectively. a.u., arbitrary units.

example, AB_4 , AB_5 , AB_6 , AB_{13}) might form when both charged and neutral nanoparticles of type B are incorporated into the structures. The presence of differently charged nanoparticles in the colloidal solutions (Fig. 2a and Supplementary Fig. 5) could also contribute to the simultaneous formation of different BNSLs. Intentional addition of a large concentration of charged species into a solution of nanoparticles might reduce the Debye screening length down to R_0 , relaxing the strict rules for BNSL charge neutrality and allowing a range of new structures to be formed¹⁰.

Tuning the charge state of the nanoparticles allows us to direct the self-assembly process. Reproducible switching between different BNSL structures has been achieved by adding small amounts of carboxylic acids, TOPO or dodecylamine to colloidal solutions of PbSe (PbS, Fe_2O_3 , and so on) and metal (Au, Ag, Pd) nanocrystals. Figure 3 demonstrates how these additives direct the formation of specific BNSL structures. Combining native solutions of 6.2 nm PbSe and 3.0 nm Pd nanoparticles (particle concentration ratio $\sim 1:5$) results in the formation of several BNSL structures with MgZn_2 and cuboctahedral AB_{13} lattices dominating. However, the same nanoparticles assemble into orthorhombic AB- and AlB_2 -type superlattices after adding oleic acid (Fig. 3a), and into NaZn_{13} - or

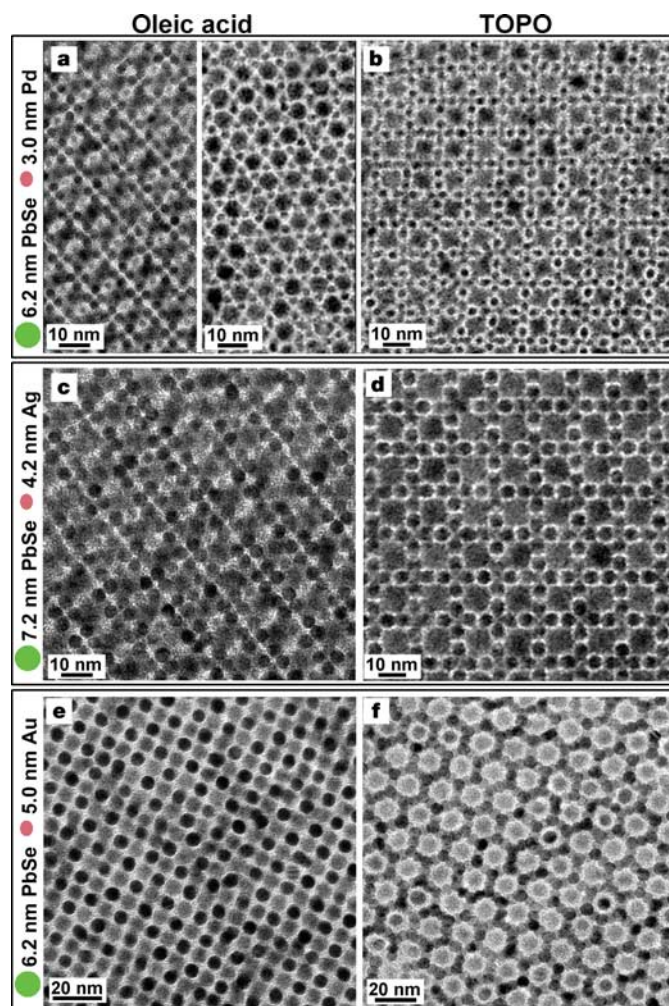


Figure 3 | TEM images of binary superlattices self-assembled in the presence of 4 mM oleic acid (left column) and 6 mM tri-*n*-octylphosphine oxide, TOPO (right column). **a**, 6.2 nm PbSe and 3.0 nm Pd nanoparticles self-assembled into orthorhombic AB- and AlB_2 -type BNSLs, and **b**, into NaZn_{13} -type BNSL. **c**, **d**, 7.2 nm PbSe and 4.2 nm Ag nanoparticles self-assembled into orthorhombic AB and cuboctahedral AB_{13} BNSLs, respectively. **e**, **f**, 6.2 nm PbSe and 5.0 nm Au nanoparticles self-assembled into CuAu-type and CaCu_5 -type BNSLs, respectively.

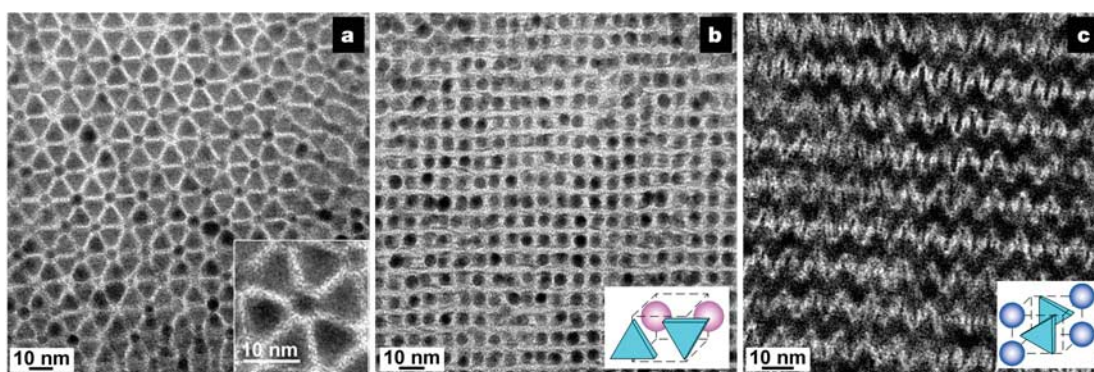


Figure 4 | TEM images and proposed unit cells of binary superlattices self-assembled from triangular nanoplates and spherical nanoparticles. **a, b**, Self-assembled from LaF_3 triangular nanoplates (9.0 nm side) and 5.0 nm Au nanoparticles; **c**, self-assembled from LaF_3 triangular nanoplates

and 6.2 nm PbSe nanocrystals. The insets show **a**, a magnified image, and **b**, **c**, proposed unit cells of the corresponding superlattices. The structure shown in **a** forms on silicon oxide surfaces, while structures shown in **b** and **c** form preferentially on amorphous carbon substrates.

cuboctahedral AB_{13} -type BNSLs after the addition of dodecylamine or TOPO, respectively (Fig. 3b). In the AB_{13} -type BNSL, metal particles assemble into icosahedral (NaZn_{13}) or cuboctahedral (cuboctahedral AB_{13}) clusters, with each large PbSe particle surrounded by 24 metal spheres at the vertices of a snub cube⁷. In the presence of TOPO the metal nanoparticles are neutral (Fig. 2f), favouring formation of the Pd_{13} (Au_{13} , Ag_{13}) clusters. The clusters of metal nanoparticles in turn provide screening of the charges on PbSe nanocrystals in the AB_{13} -type BNSL.

Surveys of many samples show that the addition of a carboxylic acid to solutions of PbSe–Pd, PbSe–Au, PbSe–Ag and PbSe– Fe_2O_3 nanoparticle mixtures results in either AB or AB_2 superlattices (Fig. 3c, e), whereas the addition of TOPO to mixtures of the same nanoparticles favours growth of AB_{13} (if $\gamma < \sim 0.65$) or AB_5 (if $\gamma > \sim 0.65$) BNSLs (Fig. 3d, f). Thus the space-filling principles and particle charging work in combination to determine the structure. Adjusting the relative concentrations of A and B particles can be used as an additional tool with which to control the BNSL structure. For example, in presence of TOPO, AB_4 BNSLs can form when the A:B ratio is $\sim 1:1$, whereas exclusively AB_{13} forms in the presence of large excess B particles.

In contrast to particles with amorphous or polycrystalline morphology, nanocrystals allow exploitation of the inherent crystal anisotropy to precisely engineer nanocrystal shape²². The nanocrystal shape can in turn be used as a powerful tool to engineer the structure of the self-assembled BNSLs. For example, Fig. 4 shows several BNSLs self-assembled from LaF_3 triangular nanoplates and spherical Au or PbSe nanocrystals. In the LaF_3 –Au system, the LaF_3 nanoplates lie flat on silicon oxide surface (Fig. 4a) and stand on edge when assembled on amorphous carbon (Fig. 4b and c), demonstrating how the choice of substrate can be used to control the BNSL structure.

It is specifically at the nanoscale that the van der Waals, electrostatic, steric repulsion and the directional dipolar interactions can contribute to the interparticle potential with comparable weight^{18,23,24,25}. These, together with the effects of particle substrate interactions and space-filling (entropic) factors, combine to determine the BNSL structure. The non-equilibrium nature of our evaporative self-assembly process adds additional complexity²⁶. Precise control of nanoparticle size, shape and composition allows us to engineer electronic, optical and magnetic properties of nanoparticle building blocks. Assembling these nanoscale building blocks into a wide range of BNSL systems provides a powerful modular approach to the design of ‘metamaterials’ with programmable physical and chemical properties.

METHODS

Nanoparticle synthesis. Au, Ag and Pd nanoparticles were prepared by modifying the method of ref. 27. Metal salts were dissolved in 10 ml of toluene with ultrasonication in the presence of dodecyldimethylammonium bromide

(DDAB). For synthesis of 5.0 nm Au and 4.2 nm Ag nanoparticles, we used 0.034 g AuCl_3 and 0.025 g AgNO_3 , respectively, and 0.0925 g DDAB. 3.0 nm Pd nanocrystals were synthesized from 0.0237 g PdCl_2 with 0.157 g DDAB. Forty microlitres of a 9.4 M aqueous solution of NaBH_4 were added drop-wise to the solution of metal salt with vigorous stirring. After 20 min, 0.8 ml 1-dodecanethiol was added and the stirring was continued for five more minutes. The nanoparticles were precipitated by adding ethanol, and the solid redispersed in 10 ml toluene in the presence of 0.8 ml 1-dodecanethiol and refluxed for 30 min under nitrogen. Fe_2O_3 nanocrystals were synthesized by methods adapted from ref. 28. Briefly, 11 nm and 13.4 nm Fe_2O_3 nanocrystals were synthesized by injecting 0.2 ml iron pentacarbonyl into 10 ml trioctylamine in the presence of 0.65 g oleic acid at 270 °C and 250 °C, respectively. After heating of the reaction mixtures at 320 °C for 1 h, the reaction mixture was cooled to room temperature. 0.17 g trimethylamine N-oxide was added to oxidize the iron nanoparticles to $\gamma\text{-Fe}_2\text{O}_3$, and the reaction mixture was heated to 130 °C for 1.5 h and 320 °C for 1 h. Details of the synthesis of PbSe, PbS and LaF_3 nanocrystals can be found in refs 7, 29 and 30, respectively.

Preparation of binary superlattices. A substrate (for example, a carbon- or silicon oxide-coated transmission electron microscope (TEM) grid, a silicon nitride membrane or an alkyl-functionalized silicon chip) was placed in a glass vial containing a colloidal solution of nanoparticles. The vial was placed tilted by 60°–70° inside a low-pressure chamber. Ordered binary assemblies formed upon evaporation of the solvent. Toluene and mixtures of toluene with tetrachloroethylene or chloroform were used as solvents ($\sim 1:1$ by volume). The best binary assemblies (as determined by the length scale of ordering and a low occurrence of defects) were obtained by evaporating relatively concentrated colloidal solutions at 45 °C under reduced pressure (~ 3.2 kPa).

Structural analysis. A Philips CM12 TEM operating at 120 kV was used to image the structure of the assemblies. Three-dimensional descriptions of the superlattices were developed by surveying large regions of the samples, to categorize all the crystal orientations, and recording a series of two-dimensional projections down the major symmetry axes. Tilting of the samples allowed observation of additional orientations not expressed in the plan view images. To assign the observed structures to crystallographic space groups, we built three-dimensional lattice models for the 180 most common space groups using Accelrys MS Modelling 3.1 software. The TEM images were compared with simulated projections to match the symmetry of our superlattices. We also performed a comparison of experimental small-angle electron diffraction patterns taken over larger areas, and the two-dimensional Fourier transformation power spectra of real space TEM images and the fast Fourier transform power spectra of the simulated projections to assure consistency.

Electrophoretic mobility measurements. These were performed by electrophoretic light scattering using a Zetasizer Nano ZS Series (Malvern), allowing measurements in non-polar organic solvents. We used chloroform solutions with nanoparticle concentrations ~ 5 times higher than those used for growing binary superlattices. The concentrations of additives (oleic acid and TOPO) were similar to those used for directing BNSL self-assembly. After preparation, the colloidal solutions were left in the dark for several hours to allow the systems to equilibrate before each measurement.

Received 20 August; accepted 2 November 2005.

- Shenton, W., Pum, D., Sleytr, U. & Mann, S. Synthesis of cadmium sulphide superlattices using self-assembled bacterial S-layers. *Nature* **389**, 585–587 (1997).

2. Guarini, K. W., Black, C. T. & Yeung, S. H. I. Optimization of diblock copolymer thin film self assembly. *Adv. Mater.* **14**, 1290–1294 (2002).
3. Redl, F. X., Cho, K.-S., Murray, C. B. & O'Brien, S. Three-dimensional binary superlattices of magnetic nanocrystals and semiconductor quantum dots. *Nature* **423**, 968–971 (2003).
4. Kiely, C. J., Fink, J., Brust, M., Bethel, D. & Schiffrin, D. J. Spontaneous ordering of bimodal ensembles of nanoscopic gold clusters. *Nature* **396**, 444–446 (1998).
5. Shevchenko, E. V. *et al.* Colloidal synthesis and self-assembly of CoPt₃ nanocrystals. *J. Am. Chem. Soc.* **124**, 11480–11485 (2002).
6. Saunders, A. E. & Korgel, B. A. Observation of an AB phase in bidisperse nanocrystal superlattices. *ChemPhysChem* **6**, 61–65 (2005).
7. Shevchenko, E. V., Talapin, D. V., O'Brien, S. & Murray, C. B. Polymorphism in AB₁₃ nanoparticle superlattices: An example of semiconductor-metal metamaterials. *J. Am. Chem. Soc.* **127**, 8741–8747 (2005).
8. Murray, M. J. & Sanders, J. V. Close-packed structures of spheres of two different sizes II. The packing densities of likely arrangements. *Phil. Mag. A* **42**, 721–740 (1980).
9. Eldridge, M. D., Madden, P. A. & Frenkel, D. Entropy-driven formation of a superlattice in a hard-sphere binary mixture. *Nature* **365**, 35–37 (1993).
10. Leunissen, M. E. *et al.* Ionic colloidal crystals of oppositely charged particles. *Nature* **437**, 235–240 (2005).
11. Bartlett, P. & Campbell, A. I. Three-dimensional binary superlattices of oppositely charged colloids. *Phys. Rev. Lett.* **95**, 128302 (2005).
12. Bolhuis, P. G., Frenkel, D., Mau, S.-C. & Huse, D. A. Entropy difference between the face-centred cubic and hexagonal close-packed crystal structures. *Nature* **388**, 235–236 (1997).
13. Pusey, P. N. & van Megen, W. Phase behaviour of concentrated suspensions of nearly hard colloidal spheres. *Nature* **320**, 340–342 (1986).
14. Wong, S., Kitaev, V. & Ozin, G. A. Colloidal crystal films: Advances in universality and perfection. *J. Am. Chem. Soc.* **125**, 15589–15598 (2003).
15. Cottin, X. & Monson, P. A. Substitutionally ordered solid solutions of hard spheres. *J. Chem. Phys.* **102**, 3354–3360 (1995).
16. Sanders, J. V. & Murray, M. J. Ordered arrangements of spheres of two different sizes in opal. *Nature* **275**, 201–203 (1978).
17. Hachisu, S. & Yoshimura, S. Optical demonstration of crystalline superstructures in binary mixtures of latex globules. *Nature* **283**, 188–189 (1980).
18. Shim, M. & Guyot-Sionnest, P. Permanent dipole moment and charges in colloidal semiconductor quantum dots. *J. Chem. Phys.* **111**, 6955–6964 (1999).
19. Krauss, T. D. & Brus, L. E. Charge, polarizability, and photoionization of single semiconductor nanocrystals. *Phys. Rev. Lett.* **83**, 4840–4843 (1999).
20. Islam, M. A. & Herman, I. P. Electrodeposition of patterned CdSe nanocrystal films using thermally charged nanocrystals. *Appl. Phys. Lett.* **80**, 3823–3825 (2002).
21. O'Brien, R. W. & White, L. R. Electrophoretic mobility of a spherical colloidal particle. *J. Chem. Soc. Farad. Trans. II* **74**, 1607–1626 (1978).
22. Yin, Y. & Alivisatos, A. P. Colloidal nanocrystal synthesis and the organic–inorganic interface. *Nature* **437**, 664–670 (2005).
23. Korgel, B. A., Fullam, S., Connolly, S. & Fitzmaurice, D. Assembly and self-organization of silver nanocrystal superlattices: Ordered “soft spheres”. *J. Phys. Chem. B* **102**, 8379–8388 (1998).
24. Cho, K.-S., Talapin, D. V., Gaschler, W. & Murray, C. B. Designing PbSe nanowires and nanorings through oriented attachment of nanoparticles. *J. Am. Chem. Soc.* **127**, 7140–7147 (2005).
25. Ohara, P. C., Leff, D. V., Heath, J. R. & Gelbart, W. M. Crystallization of opals from polydisperse nanoparticles. *Phys. Rev. Lett.* **75**, 3466–3469 (1995).
26. Rabani, E., Reichman, D. R., Geissler, P. L. & Brus, L. E. Drying-mediated self-assembly of nanoparticles. *Nature* **426**, 271–274 (2003).
27. Prasad, B. L. V., Stoeva, S. I., Sorensen, C. M. & Klabunde, K. J. Digestive ripening of thiolated gold nanoparticles: The effect of alkyl chain length. *Langmuir* **18**, 7515–7520 (2002).
28. Hyeon, T., Lee, S. S., Park, J., Chung, Y. & Na, H. B. Synthesis of highly crystalline and monodisperse maghemite nanocrystallites without a size-selection process. *J. Am. Chem. Soc.* **123**, 12798–12801 (2001).
29. Hines, M. A. & Scholes, G. D. Colloidal PbS nanocrystals with size-tunable near-infrared emission: Observation of post-synthesis self-narrowing of the particle size distribution. *Adv. Mater.* **15**, 1844–1849 (2003).
30. Zhang, Y.-W., Sun, X., Si, R., You, L.-P. & Yan, C.-H. Single-crystalline and monodisperse LaF₃ triangular nanoplates from a single-source precursor. *J. Am. Chem. Soc.* **127**, 3260–3261 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank V. Perebeinos, A. van Blaaderen, V. Crespi, I. Herman and L. E. Brus for discussions and R. L. Sandstrom for technical support. This work was partially supported by the MRSEC Program of the National Science Foundation, and by the New York State Office of Science, Technology and Academic Research (NYSTAR). S.O. is grateful for support from the DOE and an NSF CAREER award.

Author Contributions E.V.S. and D.V.T. contributed equally to this work. E.V.S. and D.V.T. carried out syntheses of nanoparticles, and E.V.S. investigated formation of binary nanoparticle superlattices. E.V.S. and D.V.T. performed modelling and structural assignment of self-assembled binary superlattices. E.V.S., D.V.T. and N.A.K. studied electrophoretic mobility of nanoparticles and worked on modelling self-assembly phenomena in binary nanoparticle colloids. S.O. and C.B.M. initiated and supervised the work. D.V.T. and C.B.M. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.V.T. (dvtalapin@lbl.gov) or C.B.M. (cbmurray@us.ibm.com).

LETTERS

Abrupt reversal in ocean overturning during the Palaeocene/Eocene warm period

Flavia Nunes¹ & Richard D. Norris¹

An exceptional analogue for the study of the causes and consequences of global warming occurs at the Palaeocene/Eocene Thermal Maximum, 55 million years ago. A rapid rise of global temperatures during this event accompanied turnovers in both marine^{1–3} and terrestrial biota⁴, as well as significant changes in ocean chemistry^{5,6} and circulation^{7,8}. Here we present evidence for an abrupt shift in deep-ocean circulation using carbon isotope records from fourteen sites. These records indicate that deep-ocean circulation patterns changed from Southern Hemisphere overturning to Northern Hemisphere overturning at the start of the Palaeocene/Eocene Thermal Maximum. This shift in the location of deep-water formation persisted for at least 40,000 years, but eventually recovered to original circulation patterns. These results corroborate climate model inferences that a shift in deep-ocean circulation would deliver relatively warmer waters to the deep sea, thus producing further warming⁹. Greenhouse conditions can thus initiate abrupt deep-ocean circulation changes in less than a few thousand years, but may have lasting effects; in this case taking 100,000 years to revert to background conditions.

Ice-age changes in insolation and surface ocean density are known to contribute to abrupt shifts in deep-ocean circulation with significant consequences for regional, even global, climate evolution. However, we have few analogues for large-scale circulation changes associated with global warming. Previous paleoceanographic work on the Palaeocene/Eocene Thermal Maximum (PETM) suggests that global warming in the earliest Eocene may have contributed to large-scale changes in deep-ocean circulation. One hypothesis proposes that lower temperature gradients between the Equator and the poles may have shifted from temperature-driven (thermohaline) to salinity-driven (halothermal) deep-ocean circulation^{7,8,10}. Other possible regime changes could have involved the subduction of thermocline waters in subtropical latitudes or a full reversal in the meridional circulation at high latitudes⁹.

Here we use carbon isotopes from 14 deep-sea sites (Fig. 1) as a nutrient tracer to reconstruct changes in deep-ocean circulation across the Palaeocene/Eocene boundary and test different models for the role of circulation in global climate. The longer a water mass is isolated from the surface, the more nutrients and ¹²C it acquires, imparting a negative $\delta^{13}\text{C}$ signature as deep waters 'age'¹¹. Therefore, carbon isotopes trace nutrient content in water masses, with the most positive isotopic ratios being indicative of sites of deep-water formation and increasingly negative $\delta^{13}\text{C}$ ratios being characteristic of old deep waters as they travel away from their sites of origin.

Here we compare $\delta^{13}\text{C}$ values from four ocean basins before, during, and after the carbon isotope excursion (CIE) associated with the PETM to monitor deep-ocean circulation. During the pre-excursion interval (Fig. 2, time slice A) the Southern Ocean and South Atlantic Ocean are characterized by the most positive $\delta^{13}\text{C}$ values in the late Palaeocene world of $+1.24 \pm 0.23\text{‰}$. By

comparison, the most negative $\delta^{13}\text{C}$ is found in the Pacific Ocean ($+0.74 \pm 0.10\text{‰}$, Fig. 2). Because no other basin has $\delta^{13}\text{C}$ values as positive as the Southern Ocean, this basin was probably the primary source of deep-water formation in the late Palaeocene, in agreement with previous studies¹². The core of the CIE (Fig. 2, time slice B) has a distinctly different $\delta^{13}\text{C}$ distribution pattern, with the Southern Hemisphere sites having the most negative values of $\delta^{13}\text{C}$ ($-1.21 \pm 0.21\text{‰}$), compared to Northern Hemisphere sites where the average $\delta^{13}\text{C}$ is $-0.50 \pm 0.17\text{‰}$ (Atlantic), $-0.52 \pm 0.05\text{‰}$ (Indian), and $-0.78 \pm 0.36\text{‰}$ (Pacific). The recovery interval of the CIE (Fig. 2, time slice C) sees a return to the pre-excursion pattern, with the Southern Ocean having isotopically more positive $\delta^{13}\text{C}$ ($+0.95 \pm 0.18\text{‰}$) compared to $+0.82 \pm 0.19\text{‰}$ in the North Atlantic, $+0.62 \pm 0.05\text{‰}$ in the Indian Ocean and $+0.67 \pm 0.06\text{‰}$ in the Pacific Ocean. In the post-excursion phase (Fig. 2, time slice D) Southern Ocean deep-ocean circulation gradually returns to the pre-excursion pattern, with the Southern Ocean $\delta^{13}\text{C}$ values being $+1.04 \pm 0.07\text{‰}$ on average, compared to $+0.68 \pm 0.07\text{‰}$ in the Northern Hemisphere basins.

Carbon isotope records suggest that a significant change in deep-ocean circulation took place across the Palaeocene/Eocene boundary (Fig. 3). The most positive $\delta^{13}\text{C}$ values occur in the Southern Hemisphere before and after the CIE, consistent with the formation of most deep water in the Southern Ocean (Fig. 3, time slices A and C). However, during the PETM itself, the Southern Ocean has more negative $\delta^{13}\text{C}$ when compared to most Northern Hemisphere sites and equatorial Pacific sites, an indication that the most significant contribution of deep water came from the Northern Hemisphere or Pacific during the CIE (Fig. 3B). Our data can most readily be explained by an interhemispheric switch in deep-water formation during the carbon isotope excursion. Indeed, these conclusions provide strong support for the results of global climate models that infer that changes in deep-ocean circulation were responsible for rapid warming of bottom waters through a switch in the locus of deep-water formation from the Southern Hemisphere to the Northern Hemisphere⁹, leading to the thermal dissociation of gas hydrates and catastrophic global warming at the time of the Palaeocene/Eocene boundary.

Our results are unlikely to be artefacts imposed by our correlation of the sites used in our compilation or our chronology. Our interpretation is critically dependent upon our inference that $\delta^{13}\text{C}$ in the core of the CIE is more positive in the Northern Hemisphere and equatorial Pacific sites than in the Southern Hemisphere sites—a conclusion that would be undercut if some sites systematically lacked the core of the CIE. Indeed, many sites do have a gap in the $\delta^{13}\text{C}$ record at the onset of the CIE, either because of carbonate dissolution (all sites other than Deep Sea Drilling Project (DSDP) Site 401 and Ocean Drilling Program (ODP) Site 690) or because of the apparent ecological absence of the benthic foraminifer, *Nuttallides truempyi*, used in our analysis (as at ODP Site 690). Our most complete records

¹Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92092-0208, USA.

are the Southern Ocean site ODP Site 690 and North Atlantic site DSDP Site 401 which both have well-preserved foraminifera through the entire CIE (despite the absence of *N. truempyi* for a short interval, there is continuous sedimentation of other species of foraminifera); notably, these sites can be well correlated in their $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ stratigraphy (Fig. 4) and show a strong reversal in benthic $\delta^{13}\text{C}$ gradients through the PETM (see Fig. 2, inset) consistent with our hypothesis of a switch in deep-water circulation. All high-resolution sites included in our compilation show a 'plateau' in $\delta^{13}\text{C}$ within the core of the CIE that suggests that $\delta^{13}\text{C}$ is not drifting from a still more negative value that is unseen owing to missing records. Indeed, most of our $\delta^{13}\text{C}$ records resemble those of highly complete terrestrial sequences, all of which display a prolonged interval of relatively constant $\delta^{13}\text{C}$ within the CIE¹³.

Although our data make a strong case for a reversal from Southern Ocean to Northern Hemisphere overturning during the PETM, it is less clear where the northern sites of deep-water formation were located. The $\delta^{13}\text{C}$ gradient between the Southern Ocean and the North Atlantic (0.70‰) is essentially the same as the gradient between the Southern Ocean and the Indian Ocean (0.69‰). Because of the closeness of the $\delta^{13}\text{C}$ values between these two Northern Hemisphere basins, it is difficult to trace precisely the flow path of deep water. The Pacific basin is represented by an equatorial site in this study, so we cannot make inferences about high

latitudes in the Pacific based on the information currently available. However, evidence from neodymium isotopes tends to rule out the deep Pacific as a major deep-water source during the PETM¹⁴, in agreement with the small $\delta^{13}\text{C}$ gradient observed between the Southern Ocean and Pacific Ocean (0.43‰). Ocean modelling suggests that saline surface waters issuing from the equatorial Tethys seaway should not be dense enough to fill the deep sea^{15,16}. The North Atlantic is a probable deep-water source, but there remains the possibility that there were multiple Northern Hemisphere sources of deep water during the CIE.

This work highlights the speed of the onset and recovery from warm climate circulation changes. In the Palaeocene/Eocene transition, deep-ocean circulation switched from a Southern Ocean source to a Northern Hemisphere source (probably to North Atlantic overturning) in <5 kyr. Recovery to the previous deep-ocean circulation pattern was gradual, however, with an intermediate circulation pattern lasting at least 200 kyr, as indicated by continued weak inter-basinal gradients after the CIE. In this warm climate mode, the oceans apparently do not revert to the pre-PETM pattern of overturning until the greenhouse gases that triggered the CIE are taken up by some combination of biological processes^{17,18} and enhanced silicate weathering¹⁹. Biological uptake of greenhouse gases probably involved a combination of storage in terrestrial biomass and soils¹⁸ and increased rates of marine export production^{17,20}, while silicate

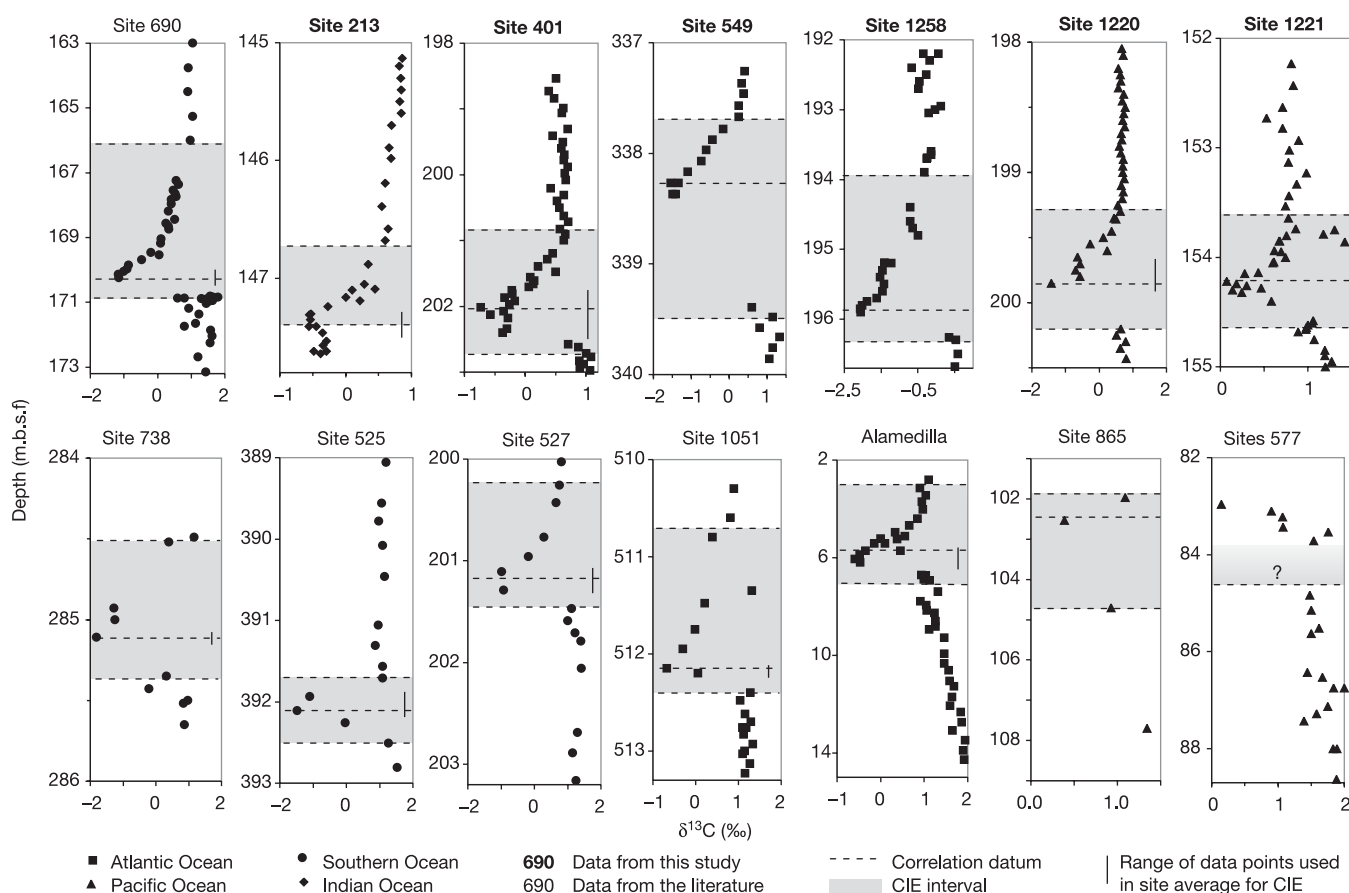


Figure 1 | Compilation of carbon isotope records. The compilation of carbon isotope records comprise a collection of eight data sets taken from the literature and six new data sets acquired at the Scripps Institution of Oceanography (See Supplementary Table 1 for data and source references). Single species analyses of $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ were made on the cosmopolitan benthic foraminifer *Nuttallides truempyi*. Deep Sea Drilling Program (DSDP) and Ocean Drilling Program (ODP) Sites 213, 401, 549, 1220, 1221 and 1258 were sampled at 5–10 cm spacing. Approximately 10–15 individuals corresponding to 50–100 μg of shell material were picked from the >150 μm fraction for single-species analyses. Carbon and oxygen

isotopic composition of these foraminifera were measured using an automated common acid bath carbonate preparation device (Fairbanks device) attached to a Finnigan MAT252 Mass Spectrometer. NBS-19 was used as a standard with seven analyses per run of 40 unknowns and all data were corrected to the Vienna Pee Dee Belemnite (VPDB) standard. The average machine error is 0.03‰ and 0.04‰ for $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ respectively. The time intervals used to calculate average isotope values for each site are: Pre-excursion, 55.234–55.334 Myr ago; CIE, indicated by black vertical bar in figure; recovery period, 54.801–55.001 Myr ago, post-excursion: 54.701–54.801 Myr ago. m.b.s.f., metres below sea floor.

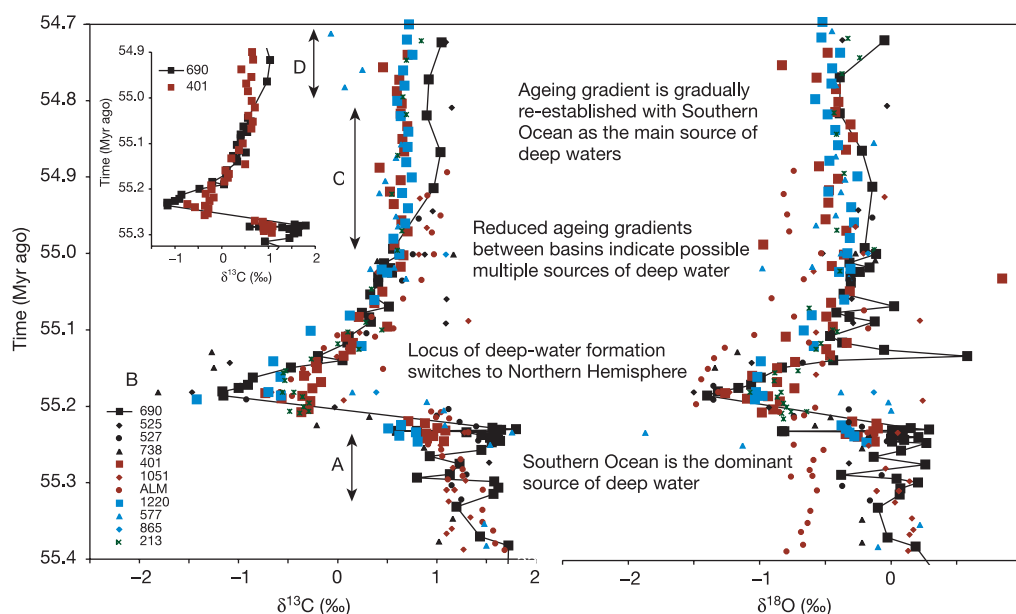


Figure 2 | Selected carbon isotope records on a common timescale. Isotope records were placed on a common age scale derived from astronomical cycle counting in sediments of ODP Site 690 (ref. 22). The absolute age for the onset of the $\delta^{13}\text{C}$ excursion was placed at 55.234 Myr ago²³. From this starting point, 21,000-yr increments were added or subtracted below or above this datum by reference to cycles in Fe and Ca (ref. 21), resulting in a well-calibrated timescale for Site 690. To apply this timescale to other records, we made the assumption that the onset of the excursion is synchronous on a global scale. Then, three tie-points were selected—one at the onset of the excursion, the minimum excursion value, and the inflection point when $\delta^{13}\text{C}$ values return to pre-excursion values. The absolute age for

the two latter tie-points, as determined through cycle counting in ODP 690, were of 55.1815 Myr ago and 55.011 Myr ago respectively. These three points were then identified in the other sites and used to calculate age scales based on linear interpolation between the tie points. We note that the use of a competing timescale for the PETM based upon cosmogenic helium fluxes²⁴ would not change our major conclusions other than to shorten the length of the 'recovery phase' (time slice C) from ~100 kyr to ~30 kyr. Otherwise, both the helium chronology and our cyclostratigraphy yield similar durations for events in the Palaeocene/Eocene boundary interval. Biostratigraphic datums were not conclusive for correlations within the $\delta^{13}\text{C}$ anomaly (see Supplementary Notes). ALM refers to the Alamedilla site, Spain.

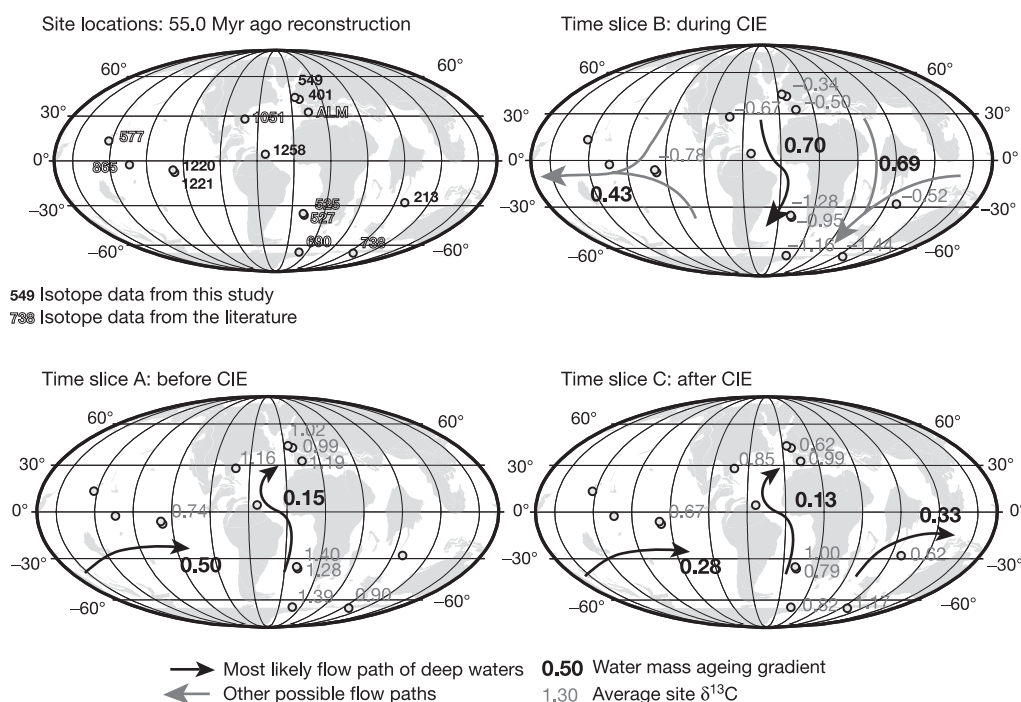


Figure 3 | Deep-ocean circulation flow paths based on carbon isotopes. Site location map and changes in $\delta^{13}\text{C}$ gradients (See Supplementary Table 2 for site and basin $\delta^{13}\text{C}$ averages). Black numbers refer to sites where new isotope data was acquired; white numbers refer to sites where isotope data is taken from the literature. Arrows indicate the flow direction of deep waters in the panels corresponding to three time slices: A, pre-excursion; B, carbon isotope excursion (CIE); and C, recovery. The Southern Ocean is the dominant source of deep water in the pre- and post-excursion periods

(A and C), whereas a switch to a Northern Hemisphere source is observed during the carbon isotope excursion (B). The recovery period (C) experiences an intermediate circulation, with small $\delta^{13}\text{C}$ ageing gradients. Data from DSDP Site 549 and ODP Sites 1221 and 1258 were omitted from this analysis because they are considered unreliable owing to geochemical overprints. Data from DSDP Sites 865 and 577 have too few data points within the PETM to produce a reliable chronology (see Supplementary Table 3).

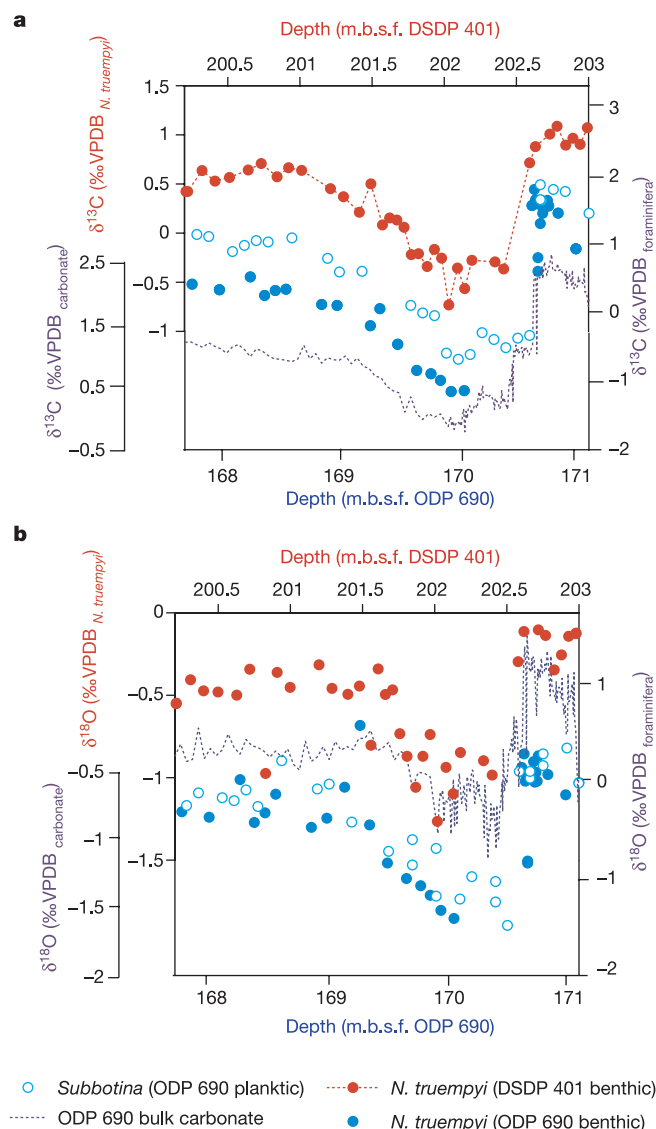


Figure 4 | Isotope stratigraphy for Southern Ocean Site ODP 690 (blue symbols) and North Atlantic Site DSDP 401 (red symbols). Comparison of $\delta^{13}\text{C}$ (a) and $\delta^{18}\text{O}$ (b) for these sites shows that both sites reproduce a series of events within the PETM, including a negative $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ event at about ~170 m in ODP Site 690. The similarity between the isotope records suggests that DSDP Site 401 is substantially complete and that the $\delta^{13}\text{C}$ gradient between the sites is robust. Indeed, neither site is likely to be missing a $\delta^{13}\text{C}$ event much more negative than seen here.

weathering was fuelled by the high surface temperatures and strong hydrologic cycle of the PETM^{13,19}. The extended recovery phase of the PETM is probably due to the ~150 kyr residence time of carbon in the exogenic carbon cycle²¹. Modern CO_2 input to the biosphere from fossil fuel sources is approaching that estimated for the PETM, raising concerns about future climate and circulation change. The PETM example suggests that anthropogenic forcing may have lasting effects not only in global climate, but in deep-ocean circulation as well.

Received 23 June; accepted 27 October 2005.

- Kelly, D. C., Bralower, T. J. & Zachos, J. C. Evolutionary consequences of the latest Paleocene thermal maximum for tropical planktonic foraminifera. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **141**, 139–161 (1998).
- Bralower, T. J. Evidence of surface water oligotrophy during the Paleocene-Eocene thermal maximum: Nannofossil assemblage data from Ocean Drilling Program Site 690, Maud Rise, Weddell Sea. *Paleoceanography* **17**, doi:10.1029/2001PA000662 (2002).
- Crouch, E. M. *et al.* Global dinoflagellate event associated with the late

- Paleocene thermal maximum. *Geology* **29**, 315–318 (2001).
- Bowen, G. J. *et al.* Mammalian dispersal at the Paleocene/Eocene boundary. *Science* **295**, 2062–2065 (2002).
- Dickens, G. R., Castillo, M. M. & Walker, J. C. G. A blast of gas in the latest Paleocene; simulating first-order effects of massive dissociation of oceanic methane hydrate. *Geology* **25**, 259–262 (1997).
- Zachos, J. C. *et al.* Rapid acidification of the ocean during the Paleocene-Eocene thermal maximum. *Science* **308**, 1611–1615 (2005).
- Kennett, J. P. & Stott, L. D. Abrupt deep-sea warming, palaeoceanographic changes and benthic extinctions at the end of the Palaeocene. *Nature* **353**, 225–229 (1991).
- Pak, D. K. & Miller, K. G. Paleocene to Eocene benthic foraminiferal isotopes and assemblages; implications for deepwater circulation. *Paleoceanography* **7**, 405–422 (1992).
- Bice, K. L. & Marotzke, J. Could changing ocean circulation have destabilized methane hydrate at the Paleocene/Eocene boundary? *Paleoceanography* **17**, 8.1–8.13 (2002).
- Kennett, J. P. *et al.* Proteus and Proto-Oceanus; ancestral Paleogene oceans as revealed from Antarctic stable isotopic results; ODP Leg 113. *Proc. ODP Sci. Res.* **113**, 865–880 (1990).
- Kroopnick, P. M. The distribution of ^{13}C of ΣCO_2 in the world oceans. *Deep-Sea Res. Part A* **32**, 57–84 (1985).
- Quilley, F., Aubry, M. P., Norris, R. D. & Berggren, W. A. Paleocene oceanography of the eastern subtropical Indian Ocean—An integrated magnetobiostratigraphic and stable isotope study of ODP Hole 761B (Wornbat Plateau). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **184**, 371–405 (2002).
- Bowen, G. J., Beerling, D. J., Koch, P. L., Zachos, J. C. & Quattlebaum, T. A humid climate state during the Palaeocene/Eocene thermal maximum. *Nature* **432**, 495–499 (2004).
- Thomas, D. J., Bralower, T. J. & Jones, C. E. Neodymium isotopic reconstruction of late Paleocene-early Eocene thermohaline circulation. *Earth Planet. Sci. Lett.* **209**, 309–322 (2003).
- Brady, E. C., DeConto, R. M. & Thompson, S. L. Deep water formation and poleward ocean heat transport in the warm climate extreme of the Cretaceous (80 Ma). *Geophys. Res. Lett.* **25**, 4205–4208 (1998).
- Bice, K. L. & Marotzke, J. Numerical evidence against reversed thermohaline circulation in the warm Paleocene/Eocene ocean. *J. Geophys. Res.* **106**, 11529–11542 (2001).
- Bains, S., Norris, R. D., Corfield, R. M. & Faul, K. L. Termination of global warmth at the Palaeocene/Eocene boundary through productivity feedback. *Nature* **407**, 171–174 (2000).
- Beerling, D. J. Increased terrestrial carbon storage across the Palaeocene-Eocene boundary. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **161**, 395–405 (2000).
- Ravizza, G., Norris, R. N., Blusztajn, J. & Aubry, M. P. An osmium isotope excursion associated with the late Paleocene thermal maximum: Evidence of intensified chemical weathering. *Paleoceanography* **16**, 155–163 (2001).
- Stoll, H. M. & Bains, S. Coccolith Sr/Ca records of productivity during the Paleocene-Eocene thermal maximum from the Weddell Sea. *Paleoceanography* **18**, doi:10.1029/2002PA000875 (2003).
- Rohl, U., Bralower, T. J., Norris, R. D. & Wefer, G. New chronology for the late Paleocene thermal maximum and its environmental implications. *Geology* **28**, 927–930 (2000).
- Norris, R. D. & Rohl, U. Carbon cycling and chronology of climate warming during the Palaeocene/Eocene transition. *Nature* **401**, 775–778 (1999).
- Wing, S. L., Bao, H. & Koch, P. L. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Wing, S. L.) 197–237 (Univ. Cambridge, Cambridge, UK, 2000).
- Farley, K. A. & Eltgroth, S. F. An alternative age model for the Paleocene-Eocene thermal maximum using extraterrestrial He-3. *Earth Planet. Sci. Lett.* **208**, 135–148 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Charles for discussions and for assistance with the mass spectrometer at SIO, and P. Worstell for assistance in the laboratory. This research used samples and data provided by the Ocean Drilling Program (ODP). ODP is sponsored by the US National Science Foundation (NSF) and participating countries under management of Joint Oceanographic Institutions (JOI), Inc. Funding for this research was provided by the National Science Foundation and the US Science Support Program (to RDN).

Author Contributions F.N. performed the data acquisition, and manuscript preparation; R.D.N. was responsible for project planning, manuscript revision and financial support. Both authors contributed equally in data analysis and interpretation.

Author Information Reprints and permissions information is available at ngp.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.N. (fnunes@ucsd.edu).

LETTERS

Three-dimensional brittle shear fracturing by tensile crack interaction

David Healy¹, Richard R. Jones² & Robert E. Holdsworth³

Faults in brittle rock are shear fractures formed through the interaction and coalescence of many tensile microcracks^{1–3}. The geometry of these microcracks and their surrounding elastic stress fields control the orientation of the final shear fracture surfaces^{3,4}. The classic Coulomb–Mohr failure criterion⁵ predicts the development of two conjugate (bimodal) shear planes that are inclined at an acute angle to the axis of maximum compressive stress. This criterion, however, is incapable of explaining the three-dimensional polymodal fault patterns that are widely observed in rocks^{6–9}. Here we show that the elastic stress around tensile microcracks in three dimensions promotes a mutual interaction that produces brittle shear planes oriented obliquely to the remote principal stresses, and can therefore account for observed polymodal fault patterns. Our microcrack interaction model is based on the three-dimensional solution of Eshelby^{10,11}, unlike previous models^{4,12} that employed two-dimensional approximations. Our model predicts that shear fractures formed by the coalescence of interacting mode I cracks will be inclined at a maximum of 26° to the axes of remote maximum and intermediate compression. An improved understanding of brittle shear failure in three dimensions has important implications for earthquake seismology and rock-mass stability, as well as fluid migration in fractured rocks¹³.

The classical andersonian model of faulting¹⁴, based on the Coulomb–Mohr failure criterion⁵, can explain conjugate (bimodal) sets of faults (Fig. 1a) in the brittle regime. According to the Coulomb–Mohr criterion, shear failure occurs along surfaces on which the ratio of shear stress to normal stress is a maximum. Stress calculations show that this condition is satisfied by two planes inclined at acute angles to the axis of maximum compressive stress σ_1 (Fig. 1a) and parallel to the axis of the intermediate stress σ_2 . Rock deformation experiments¹ seem to confirm the validity of this criterion, in terms of both the stress magnitude during failure and the orientation of faults with respect to σ_1 . However, as most of these experiments are conducted using a cylindrical geometry, with an axially applied load (σ_1) greater than a radial load ($\sigma_2 = \sigma_3$) applied through a confining jacket, they cannot provide relevant information on the orientations of shear fractures in a three-dimensional (3D) stress state where $\sigma_1 > \sigma_2 > \sigma_3$.

Field and experimental observations indicate the insufficiency of the Coulomb–Mohr criterion. Reches and Dieterich⁹ performed true triaxial (multi- or polyaxial) experiments with $\sigma_1 > \sigma_2 > \sigma_3$, and measured the orientation of the fault surfaces with respect to the axes of the three principal stresses and the magnitudes of principal strain rates. Their experiments⁹ generated polymodal shear fractures oriented at varying degrees of obliquity to the applied stresses in various rock types. Field observations of contemporaneous polymodal fracture patterns include examples from sedimentary^{7,15}, volcanic¹⁶ and metamorphic basement¹⁷ rocks. These natural and experimental examples show that failure criteria that predict only

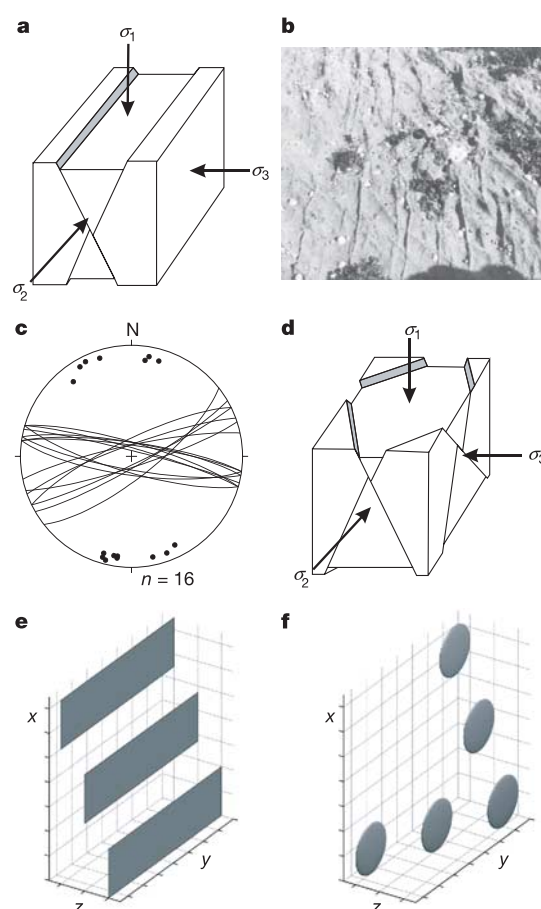


Figure 1 | Contrasting patterns of conjugate (bimodal) and polymodal faults. **a**, Conjugate faults form symmetrically about the remote σ_1 direction, which bisects the acute angle between the planes, and intersect along a line parallel to the remote σ_2 direction. **b**, Polymodal faults cutting sandstone at Gruinard Bay in northwest Scotland. A rhombohedral pattern is clear, although in detail, some fractures anastomose between definite limits. **c**, Stereonet (lower-hemisphere equal-area) of poles to fracture planes shown in **b**. N is the north azimuth and n is the number of measurements. The poles are clustered and show a quadrimodal pattern^{6–9}. **d**, Polymodal faults intersect to form rhombohedral traces on outcrop surfaces with σ_1 and σ_2 bisecting the acute angles between the failure planes. **e**, Crack configuration used by ref. 4, where each crack is very thin in the z direction and infinite in the y direction. **f**, Crack configuration used in this study, with each crack modelled as a finite oblate spheroid. All cracks are aligned with their circular equatorial planes parallel to the x – y plane.

¹Rock Deformation Laboratory, Department of Earth and Ocean Sciences, University of Liverpool, Liverpool L69 3GP, UK. ²e-Science Research Institute, ³Reactivation Research Group, Department of Earth Sciences, University of Durham, Durham DH1 3LE, UK.

conjugate bimodal faults are at best an approximation. Andersonian conjugate faults (Fig. 1a) can accommodate only a two-dimensional (2D) plane strain, whereas displacement on polymodal fault sets (Fig. 1d) produces 3D strains, which is the general deformation in the Earth^{18,19}. Furthermore, a more complete and general model is needed to improve the understanding of brittle failure in 3D.

Experiments^{1,2,20} and theoretical derivations^{3,4} have led to the understanding that brittle shear fractures in rocks nucleate and propagate through the interaction and coalescence of sub-parallel tensile microcracks. Acoustic emission observations in rock deformation experiments² suggest that tensile microcrack arrays are localized in the region of the final shear fracture, and that coalescence of the interacting microcracks into a through-going shear fracture is catastrophic and occurs just before sample failure²¹. Previously modelled interactions among microcracks were restricted to a 2D approximation^{4,12} and assumed that mode I tensile cracks have infinite length parallel to σ_2 (y axis in Fig. 1e). Our numerical models, on the other hand, consider each tensile microcrack as a finite 3D oblate ellipsoidal void (Fig. 1f). We retain the concept of a homogeneous, isotropic linear elastic matrix, and focus on the determination of the 3D distribution of elastic stress around interacting cracks to explain the formation of shear fractures in 3D.

To calculate the elastic field around a crack in 3D we assume that the cracks are ellipsoidal voids, and we employ the Eshelby^{10,11,22} solution for a 'penny-shaped' void. We set the elastic properties of the void to zero to approximate a fluid-filled microcrack (Fig. 2a). In this linear elastic model we present only the relative values of crack dimensions and stresses. The elastic stress field around an isolated tensile mode I crack (Fig. 2b) is subjected to a uniaxial tensile stress, σ_{zz} (or equivalent uniform internal pressure⁴). We do not impose a remote compressive stress field and it is assumed that tensile microcracks in a rock subjected to a triaxially compressive stress field will in general be normal to the axis of remote least compression (that is parallel to the x - y plane in Fig. 2)^{4,20,23}. The model prescribes the location and geometry of the interacting tensile microcracks, and the elastostatic solution does not incorporate either microcrack nucleation or propagation. We exclude these processes to simplify the model, and to focus on the geometry of tensile crack interaction.

The form of the modelled stress distribution (Fig. 2c) is similar to previous results for a mode I crack under uniaxial tension^{4,24}, with bi-lobate regions of increased tensile stress at both sides of the crack. Owing to the symmetry, the stress field in the y - z plane is identical to that in the x - z plane. This implies that the same stress interaction with neighbouring cracks applies in the x - z and y - z planes (Fig. 2d), and in 3D space. In 3D, the region of increased tensile stress around a tensile microcrack forms an indented ring-shaped, toroidal zone enclosing the crack and extending out to either side (Fig. 2e).

Using a 3D distribution of tensile stress around a single tensile microcrack (Fig. 2e), we now assess the nature of the interaction among a pair of neighbouring cracks. Following ref. 4, we assume that the key factor governing interaction is the magnitude of the crack-normal tensile stress. The lobes of increased crack-normal tensile stress for each crack merge and interact with those of the adjacent crack, and the stress at the ends of any pair is increased relative to that of a single crack⁴. This mutual amplification forms the basis of the self-organizing nucleation and runaway propagation of the shear fracture nucleus⁴. The average crack-normal stress acting on a neighbouring crack is:

$$\text{CNS}_{\text{average}} = \int_{-L/2}^{+L/2} \sigma_{zz} dr$$

where $L = 2a$ ($= 2b$ for 'penny-shaped' cracks), and r is the radial distance from the crack tip. We compute $\text{CNS}_{\text{average}}$ by numerical integration and then calculate the locus of the maximum of $\text{CNS}_{\text{average}}$ around the crack. This locus forms a hyperbola in 2D with tangents symmetrical about the x (or y) axis (Fig. 3a). The angle

θ these tangents make with any crack-parallel axis is a constant of the elastic stress field around an ellipsoidal crack, and as shown in the graph in Fig. 3a, is approximately 26° . This is close to the value of $\pm 34^\circ$ reported by ref. 4 from their 2D model and we ascribe the discrepancy to the differences between 2D and 3D formulations of the stress field and a small component of error from inexact numerical calculation. In Fig. 3a, a neighbouring crack located at A falls on the locus of maximum tensile interaction and may coalesce with the lower crack and form a through-going shear fracture inclined to the vertical at an angle less than θ . A similar crack at B will interact to form a shear fracture inclined at angle θ . Irrespective of the spacing of the interacting cracks, the envelope to the locus of

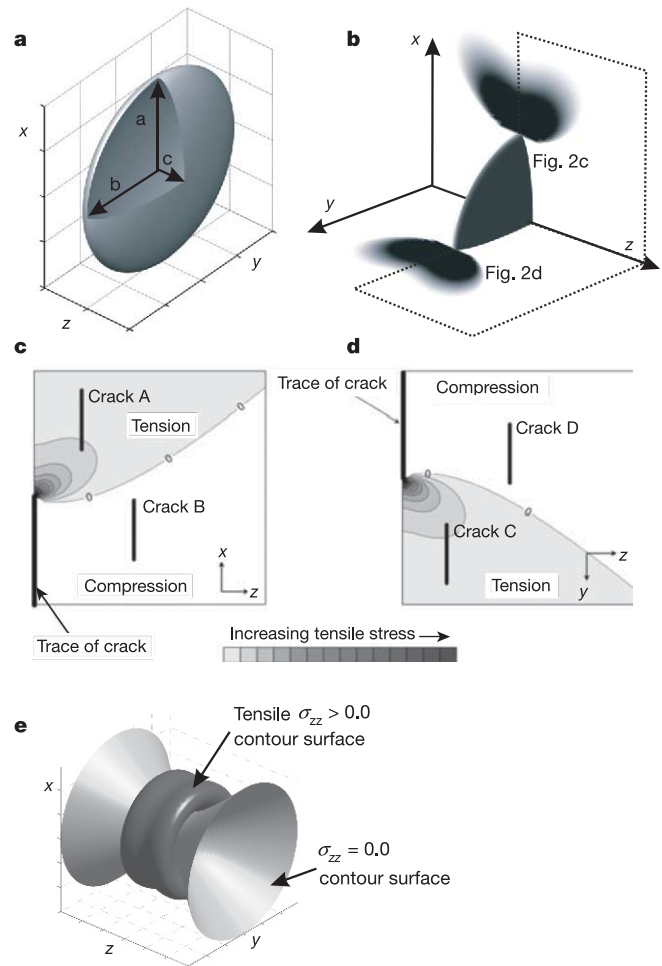


Figure 2 | Elastic stress field around an isolated 'penny-shaped' crack.

a, Reference frame used in this study with a 'penny-shaped' tensile microcrack modelled as an oblate spheroidal void with semi-axes $a = b \gg c$, and aspect ratio $a/c = 100$. Semi-axes a , b and c are aligned with the x , y and z coordinate axes respectively. **b**, 3D plot to show the elastic stress field around an isolated tensile crack^{10,11,22}. The mapped value is the σ_{zz} component of the perturbed stress field, which is the most important when considering tensile interactions of cracks oriented with their short c axes parallel to z . Regions with compressive stress appear white; increasing tensile stress is shown as darker grey. **c**, Region outlined in **b**. Neighbouring crack A lies in the tensile field while crack B lies in the compressive field. Crack A will tend to be opened and crack B will tend to be closed by the elastic field of the parent crack. **d**, Region outlined in **b**. Neighbouring crack C lies in the tensile field of the parent crack, while crack D lies in the compressive field. Crack C will tend to be opened and crack D will tend to be closed by the elastic field of the parent crack. **e**, 3D view of the isolated crack and the surrounding σ_{zz} field. The elastic stress field of an isolated crack will promote *en echelon* tensile interaction with neighbouring cracks in 3D, and this interaction is not confined to a single (for example, x - z) plane.

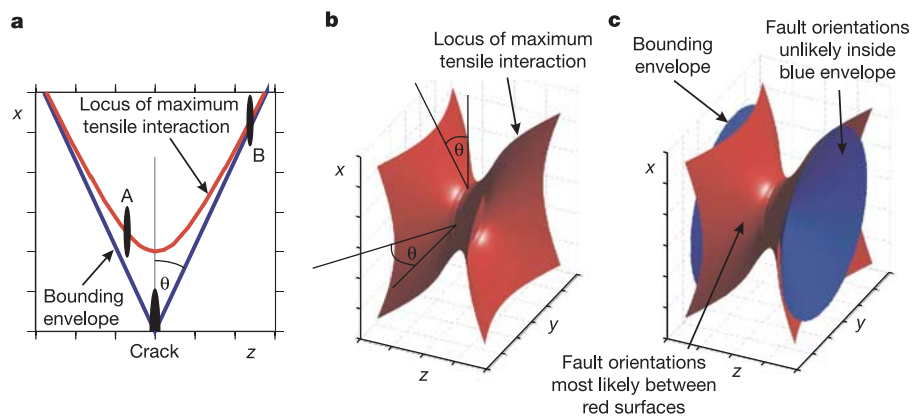


Figure 3 | Locus of maximum tensile interaction. **a**, 2D plot in the x - z plane showing the hyperbolic locus of maximum tensile interaction (red) and corresponding bounding envelope (blue). The angle θ measures the inclination of the bounding envelope from the plane of the crack, and defines an upper limit to the orientation of likely shear fractures formed by interaction and coalescence among neighbouring cracks. **b**, 3D view of the

locus of maximum tensile interaction, now shown as a hyperboloid with radial symmetry (for a 'penny-shaped' crack) about the z axis. **c**, 3D view of the conical (blue) tangential envelopes to the locus of maximum tensile interaction (red). The most likely shear failure planes are expected to lie within the volume between the outer surfaces of the cones.

maximum tensile interaction provides a limiting angle controlling the relative position of cracks that are most likely to preferentially interact and coalesce.

Assuming that shear fractures nucleate through the mutual interaction of neighbouring tensile cracks, shear planes are therefore much less likely to develop at θ angles greater than 26° . In 3D the locus of maximum tensile interaction forms a hyperboloid (Fig. 3b) with two conical tangential envelopes symmetrical about the z axis (Fig. 3c). These cones can be considered as the rotation of the tangents (in Fig. 3a) through 360° about the z axis. Shear fracture planes formed by tensile crack coalescence are equally likely to form in any orientation bounded by the outer surface of these cones (Fig. 3c).

The conical bounding envelopes for predicted shear fracture orientation trace out small circles on a stereonet (Fig. 4a). Brittle shear fracture data measured in the field at three separate stations are shown in Fig. 4a–c. The bounding cones delimit zones of likely fracture plane orientation. The fit between model predictions and field data are good, and, in the laboratory results of ref. 9 (their figure 5), the fracture data are limited by a bounding envelope of just over 25° . The field data in Fig. 4c cluster into four groups that show a quadrimodal pattern with orthorhombic symmetry^{6,7}. This higher-order symmetry is not predicted by our model, but is explained by the slip model of ref. 8.

Our future work will model the interactions of cracks of general

ellipsoidal shape ($a > b > c$), rather than the spheroidal shapes ($a = b \gg c$) used in this study. We suspect that the orthorhombic symmetry of the stress field associated with true ellipsoidal cracks may impart significant changes to the 3D form of the locus of maximum tensile interaction, as might the remote stress (strain) state or elastic anisotropy in the rock. These factors, absent in the current model, may promote the development of more ordered fracture sets such as quadrimodal arrays (Fig. 4c). We also note that natural and experimentally produced⁹ polymodal fault patterns often contain anastomosing and curvi-planar fracture surfaces. This variation in the orientation of macroscopic shear surfaces (Fig. 4a, b) may be controlled by mutually interacting microcracks preferentially located on the curvi-planar locus of maximum tensile interaction (Fig. 3b). Work in progress using digital methods of detailed field survey²⁵ aims to quantify the degree and extent of fault plane curvature in natural examples. The small circle girdle distributions of the poles to shear fracture planes resemble those described by ref. 26 in their slip-tendency analysis (their figure 2). However, they considered the case of rocks populated with many cohesionless pre-existing fractures and found the preferred orientations of slip under a given remote stress, whereas our model predicts the formation of new shear fractures within intact rock.

Stress calculations based on linear elastic fracture mechanics (LEFM) suggest that shear fractures cannot propagate in their own plane²⁷. However, for composite shear fractures formed through the interaction and coalescence of many tensile cracks, the simplified geometry of a single shear plane used in LEFM models is not a valid approximation. The propagation of a composite macroscopic shear fracture might never be strictly in-plane to any of the constituent microcracks, but when viewed at a larger scale, the macroscopic shear failure surface tends to propagate in-plane^{4,27}. This concept is supported by acoustic emission data from rock deformation experiments² where crescent-shaped clouds of emissions representing tensile microcracking events surround the edges of the propagating shear plane. Further experimental evidence for in-plane propagation of shear fractures is provided by microstructural observations²⁸ where torsion experiments on granite produced a shear plane that propagated through obliquely oriented tensile cracks. Therefore, we believe that a brittle shear fracture nucleus, formed through the interaction and coalescence of tensile microcracks and oriented oblique to the remote principal stresses, is likely to propagate within its own plane and maintain its oblique orientation.

Our 3D model of brittle shear failure has important implications for earthquake seismology, rock-mass stability and fluid migration in

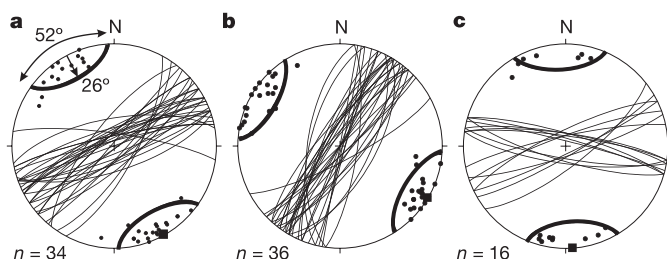


Figure 4 | Comparison of model predictions with fault data from Triassic sandstones around Grunard Bay. All plots show lower-hemisphere equal-area stereonet projections with fracture planes plotted as poles and great circles. The square symbol marks the pole to the mean plane. The conical envelopes have been plotted as small circle girdles (bold arcs) about the mean fracture direction, with apical angle 52° ($2 \times 26^\circ$) and a dip range of 90° – 64° —that is, a maximum deviation from the vertical of 26° . The faults were measured at Laide waterfall (**a**), Laide jetty (**b**) and Udrigle (**c**).

fractured rocks, because most published studies in these areas assume andersonian fault behaviour. Widely observed polymodal fracture patterns^{6–9,15–17} accommodate 3D strains with important consequences for rock properties, including strength and permeability. Our model provides a 3D micromechanical explanation for polymodal fracture patterns and improved predictions of shear fracture geometry.

Received 22 July; accepted 18 October 2005.

- Paterson, M. S. *Experimental Rock Deformation—The Brittle Field* Ch. 3 & 4 (Springer, Berlin, 1978).
- Lockner, D. A., Byerlee, J. D., Kukusenko, V., Ponomarev, A. & Sidorin, A. Observations of quasistatic fault growth from acoustic emissions. in *Fault Mechanics and Transport Properties of Rocks* (eds Evans, B. & Wong, T.) 3–31 (Academic, London, 1992).
- Scholz, C. H. *The Mechanics of Earthquakes and Faulting* Ch. 1 & 3 (Cambridge Univ. Press, Cambridge, UK, 2002).
- Reches, Z. & Lockner, D. A. Nucleation and growth of faults in brittle rocks. *J. Geophys. Res.* **99**, 18159–18173 (1994).
- Jaeger, J. & Cook, N. G. W. *Fundamentals of Rock Mechanics*, 3rd edn, Ch. 4 & 6 (Chapman & Hall, London, 1979).
- Reches, Z. Analysis of faulting in three-dimensional strain fields. *Tectonophysics* **47**, 109–129 (1978).
- Aydin, A. & Reches, Z. Number and orientation of fault sets in the field and in experiments. *Geology* **10**, 107–112 (1982).
- Reches, Z. Faulting of rocks in three-dimensional strain fields. II. Theoretical analysis. *Tectonophysics* **95**, 133–156 (1983).
- Reches, Z. & Dieterich, J. Faulting of rocks in three-dimensional strain fields. I. Failure of rocks in polyaxial, servo-control experiments. *Tectonophysics* **95**, 111–132 (1983).
- Eshelby, J. D. The determination of the elastic field of an ellipsoidal inclusion, and related problems. *Proc. R. Soc. Lond. A* **241**, 376–396 (1957).
- Eshelby, J. D. The elastic field outside an ellipsoidal inclusion. *Proc. R. Soc. Lond. A* **252**, 561–569 (1959).
- Du, Y. & Aydin, A. Interaction of multiple cracks and formation of echelon crack arrays. *Int. J. Num. Anal. Methods Geomech.* **15**, 205–218 (1991).
- Sternlof, K. R., Chapin, J. R., Pollard, D. D. & Durlafsky, L. J. Permeability effects of deformation band arrays in sandstone. *AAPG Bull.* **88**, 1315–1329 (2004).
- Anderson, E. M. *The Dynamics Of Faulting And Dyke Formation With Applications To Britain* (Oliver and Boyd, Edinburgh, 1942).
- De Paola, N., Holdsworth, R. E. & McCaffrey, K. J. W. The influence of lithology and pre-existing structures on reservoir-scale faulting patterns in transtensional rift zones. *J. Geol. Soc. Lond.* **162**, 471–480 (2005).
- Crider, J. G. Oblique slip and the geometry of normal-fault linkage: mechanics and a case study from the Basin and Range in Oregon. *J. Struct. Geol.* **23**, 1997–2009 (2001).
- Beacom, L. E., Anderson, T. B. & Holdsworth, R. E. Using basement-hosted clastic dykes as syn-rifting palaeostress indicators: an example from the basal Stoer group, northwest Scotland. *Geol. Mag.* **136**, 301–310 (1999).
- Dewey, J. F., Holdsworth, R. E. & Strachan, R. A. in *Continental Transpressional and Transtensional Tectonics* (eds Holdsworth, R. E., Strachan, R. A. & Dewey, J. F.) Vol. **135** 1–14 (Geological Society Special Publications, London, 1998).
- Jones, R. R., Holdsworth, R. E., McCaffrey, K. J. W., Clegg, P. & Tavarnelli, E. Scale dependence, strain compatibility and heterogeneity of three-dimensional deformation during mountain building: a discussion. *J. Struct. Geol.* **27**, 1190–1204 (2005).
- Moore, D. E. & Lockner, D. A. The role of microcracking in shear-fracture propagation in granite. *J. Struct. Geol.* **17**, 95–114 (1995).
- Glover, P. W. J., Gomez, J. B. & Meredith, P. G. Fracturing in saturated rocks undergoing triaxial deformation using complex electrical conductivity measurements: experimental study. *Earth Planet. Sci. Lett.* **183**, 201–213 (2000).
- Ju, J. W. & Sun, L. Z. Effective elastoplastic behaviour of metal matrix composites containing randomly located aligned spheroidal inhomogeneities. Part 1: micromechanics based formulation. *Int. J. Solids Struct.* **38**, 183–201 (2001).
- Peng, S. & Johnson, A. M. Crack growth and faulting in cylindrical specimens of Chelmsford granite. *Int. J. Rock Mech. Mining Sci.* **9**, 37–86 (1972).
- Pollard, D. D. & Segall, P. Theoretical displacements and stresses near fractures in rock: with applications to faults, joints, veins, dikes, and solution surfaces. in *Fracture Mechanics of Rock* (ed. Atkinson, B. K.) 277–349 (Academic, London, 1987).
- McCaffrey, K. J. W. *et al.* Unlocking the spatial dimension: digital technologies and the future of geoscience fieldwork. *J. Geol. Soc. Lond.* **162**, 927–938 (2005).
- Morris, A., Ferrill, D. A. & Henderson, D. B. Slip-tendency analysis and fault reactivation. *Geology* **24**, 275–278 (1996).
- Petit, J.-P. & Barquins, M. Can natural faults propagate under mode II conditions? *Tectonics* **7**, 1243–1256 (1988).
- Cox, S. J. D. & Scholz, C. H. On the formation and growth of faults: an experimental study. *J. Struct. Geol.* **10**, 413–430 (1988).

Acknowledgements We thank L. Sun for help with benchmarking the external field solution of Eshelby, and Z. Reches for constructive reviews. Thanks also to N. Kusznir for help with conference funds and P. Meredith, N. de Paola and D. Faulkner for discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.H. (dhealy@liverpool.ac.uk).

LETTERS

Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*

S. T. Dyhrman¹, P. D. Chappell², S. T. Haley¹, J. W. Moffett², E. D. Orchard¹, J. B. Waterbury¹ & E. A. Webb¹

The factors that control the growth and nitrogen fixation rates of marine diazotrophs such as *Trichodesmium* have been intensively studied because of the role that these processes have in the global cycling of carbon and nitrogen, and in the sequestration of carbon to the deep sea. Because the phosphate concentrations of many ocean gyres are low¹, the bioavailability of the larger, chemically heterogeneous pool of dissolved organic phosphorus could markedly influence *Trichodesmium* physiology. Here we describe the induction, by phosphorus stress, of genes from the *Trichodesmium erythraeum* IMS101 genome that are predicted to encode proteins associated with the high-affinity transport and hydrolysis of phosphonate compounds by a carbon–phosphorus lyase pathway. We show the importance of these genes through expression analyses with *T. erythraeum* from the Sargasso Sea. Phosphonates are known to be present in oligotrophic marine systems, but have not previously been considered to be bioavailable to marine diazotrophs. The apparent absence of genes encoding a carbon–phosphorus lyase pathway in the other marine cyanobacterial genomes suggests that, relative to other phytoplankton, *Trichodesmium* is uniquely adapted for scavenging phosphorus from organic sources. This adaptation may help to explain the prevalence of *Trichodesmium* in low phosphate, oligotrophic systems.

Marine phytoplankton account for roughly half of the annual global fixation of atmospheric CO₂ (ref. 2). In the central oceanic gyres, significant components of the phytoplankton are cyanobacteria from the genera *Synechococcus*, *Prochlorococcus*, *Crocospaera* and *Trichodesmium*. In addition to their capacity to fix CO₂, both *Crocospaera* and *Trichodesmium* are diazotrophic and therefore represent a considerable source of newly fixed nitrogen to the euphotic zones of the tropical and subtropical regimes where they occur^{3,4}. Studies from the North Atlantic show that *Trichodesmium* introduces the largest fraction of new nitrogen to the euphotic zone (~30 mg N m⁻² d⁻¹), a value that exceeds the estimated flux of nitrate across the thermocline⁵. With its high abundance and dual roles in CO₂ and N₂ fixation, *Trichodesmium* has a profound influence on the carbon and nitrogen cycles.

One of the constraints on CO₂ and N₂ fixation by *Trichodesmium* is phosphorus bioavailability⁶. Whereas dissolved inorganic phosphate (DIP) is directly available to cyanobacteria, the degree to which surface-adsorbed phosphorus (ref. 7) and the complex constituents of the dissolved organic phosphorus (DOP) pool are bioavailable is poorly understood. In the surface of oligotrophic waters, such as the North Atlantic, DOP often comprises a significant proportion of the total dissolved phosphorus (TDP)⁸. Owing to constraints in the sensitivity of ³¹P-NMR, little is known about the bond classes present in low molecular mass DOP. However, two dominant bond classes of ultrafiltered high molecular mass DOP in the ocean are monophosphate esters (C–O–P bond) and phosphonates (C–P bond)^{9,10}. Many cyanobacteria, including *Trichodesmium*, can hydrolyse monophos-

phate esters^{11,12}, enabling them to access ~75% of the ultrafiltered DOP in both the Pacific and the Atlantic^{9,10}. The other principal bond class of ultrafiltered DOP, the phosphonates, constitutes ~25% of the DOP and has the same oceanic distribution as monophosphate esters, suggesting that this bond class may also be bioavailable. However, phosphonates make up only a few per cent of the total phosphorus in phytoplankton and sinking marine particles, but accumulate to 25% in the DOP^{9,10,13}. As a result, phosphonates with their more stable C–P bond, are typically considered refractory relative to monophosphate esters in marine systems, and the bioavailability of marine phosphonates to cyanobacteria is only now being considered^{14,15}.

Two major mechanisms have been described for the utilization of phosphonates by bacteria: systems using a phosphonatase, and systems using a C–P lyase¹⁶. Phosphonatase enzymes hydrolyse specific substrates, for example, the enzyme phosphonoacetylaldehyde hydrolase (EC 3.11.1.1) hydrolyses 2-aminoethylphosphonic acid, whereas the C–P lyase pathway mediates the hydrolysis of a broad range of substrates¹⁷. In *Pseudomonas stutzeri* and *Escherichia coli*, phosphonate utilization is mediated by a cluster of 14 genes (*phnC* to *phnP*) encoding a C–P lyase pathway¹⁸. The expression of these genes is under the control of the *pho* regulon and regulated by phosphate availability¹⁷. Gene products involved in phosphonate transport are encoded by *phnC* to *phnE*. The genes *phnG* to *phnM* are thought to encode a membrane-associated C–P lyase complex, whereas several other genes, including *phnF*, *phnN*, *phnO* and *phnP*, are not required for phosphonate utilization but may be accessory proteins for the C–P lyase or transcriptional regulators¹⁷.

The genome of *T. erythraeum* IMS101 contains a gene cluster that is predicted to encode the transport and hydrolysis of phosphonate compounds by a similar C–P lyase pathway. Specifically, it has orthologues to *phnC* to *phnM*, and lacks orthologues to the accessory genes *phnF*, *phnN*, *phnO* and *phnP* (Fig. 1). Using primers designed to the *T. erythraeum* genes, we confirmed the presence of *phnD* and *phnJ* orthologues in other cultured *Trichodesmium* species (*T. tenue*, *T. thiebautii*, and *T. spiralis*)¹⁹ and found that gene sequence identities ranged between 96 and 97% over the 147-bp *phnD* and 280-bp *phnJ* amplicons (Supplementary Figs 1 and 2). These data are consistent with results of ongoing sequencing of other phosphate-regulated genes such as *pstS* and *phoA* that also have high sequence identity among *Trichodesmium* species (E.D.O., E.A.W., J.B.W. and S.T.D., unpublished results). In brief, the genes required for phosphonate transport and hydrolysis by a C–P lyase seem to be present in all of the *Trichodesmium* species tested so far.

Of the nine marine cyanobacterial genomes currently available (*Prochlorococcus* MED4, MIT9313, MIT9312 and SS120; *Synechococcus* WH8102, CC9605 and CC9902; *Crocospaera watsonii* WH8501; and *T. erythraeum* IMS101), *T. erythraeum* IMS101 is the only one that encodes the C–P lyase pathway for phosphonate utilization. The

¹Biology Department, and ²Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.

only other cyanobacterial genome that does contain the C-P lyase gene cluster is freshwater *Nostoc* PCC 7120. Much evidence suggests that the gene cluster in *T. erythraeum* IMS101 was acquired through a horizontal gene transfer event. Inspection of the genomic neighbourhood of the *T. erythraeum* IMS101 *phn* gene orthologues shows that the cluster is more similar, in both arrangement and identity, to that of *Thiobacillus denitrificans*, a sulphur-oxidizing obligate chemolithoautotrophic β -proteobacterium, than to that of *Nostoc* PCC 7102 (Fig. 1 and Supplementary Table 1). In addition, the GC content of the *Trichodesmium* *phn* genes are substantially increased (~42%) relative to the rest of the genome (34%; Fig. 1).

Phylogenetic analyses of the PhnJ protein identified three main groupings: one containing sequences from γ -proteobacteria; another consisting completely of α -proteobacteria; and a third group

consisting of sequences from diverse bacterial lineages, including the PhnJ protein from *T. erythraeum* IMS101 (Fig. 1). These data and the presence of the genes in the other species of *Trichodesmium* are consistent with the hypothesis that the *Trichodesmium* *phn* cluster was horizontally acquired before the radiation of the *Trichodesmium* genus. Furthermore, the apparent absence of the *phn* genes in the other marine cyanobacterial genomes (including *C. watsonii* WH8501, the only other oceanic diazotroph with a genome available), suggests that *Trichodesmium* could have a competitive advantage with respect to DOP utilization in the many oligotrophic regimes where they coexist. These results help to explain the global abundances of *Trichodesmium* in many low phosphate environments, and also underscore the possibility that phosphonate metabolism could represent a crucial niche adaptation by *Trichodesmium*.

To explore the role of the *phn* cluster in the metabolism of *Trichodesmium*, the expression of the *phnD* and *phnJ* genes were analysed in cells grown in different conditions. Both genes were expressed in phosphate-deficient axenic cultures of *T. erythraeum* IMS101 and were not detected in phosphate-replete cultures or iron-deficient cultures, showing that the expression of these genes is controlled by phosphate supply and not through a general stress response (Fig. 2). This is consistent with the phosphate regulation of *phn* genes observed in other microbes¹⁷, and these data highlight the importance of phosphonate utilization as a DOP acquisition strategy during periods of *Trichodesmium* phosphorus stress.

We examined field populations of *T. erythraeum* for the expression of the *phnD* and *phnJ* genes in an oligotrophic region of the western

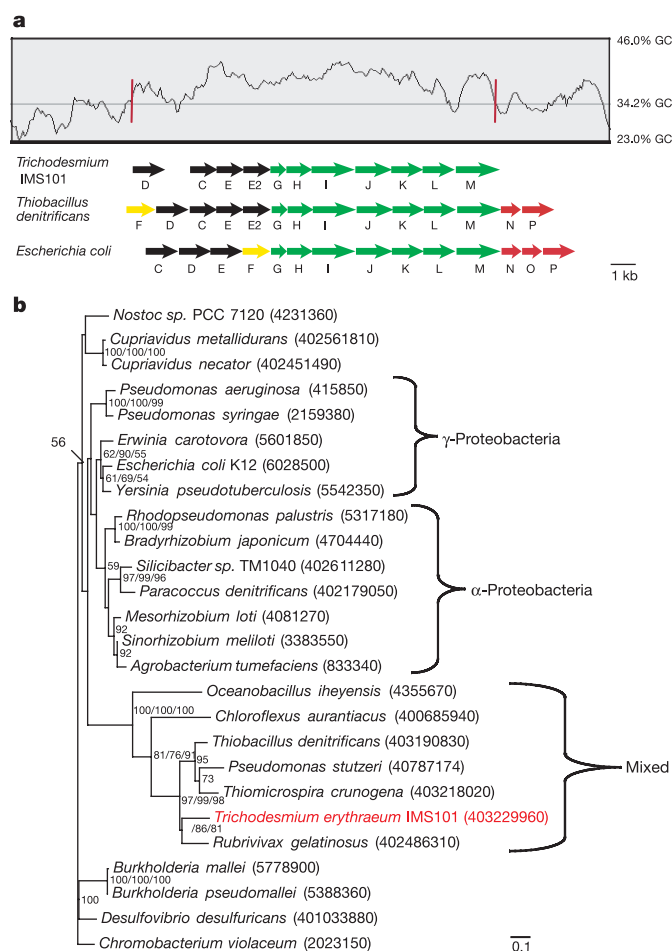


Figure 1 | DNA topology of the *phn* cluster and phylogenetic analyses of the PhnJ protein. **a**, GC content of the *phn* cluster from the *T. erythraeum* IMS101 genome. The 5' end of *phnD* and the 3' end of *phnM* are indicated by red vertical lines; the genomic average GC content (34.2%) is indicated by a black horizontal line. Also shown is the organization of the *phn* gene cluster in *T. erythraeum* IMS101 as compared with *Thiobacillus denitrificans* and *E. coli* K12. Genes encoding phosphonate transport (black) and regulation (yellow), the C-P lyase subunits (green) and accessory proteins (red) are shown. **b**, Phylogenetic relationship of the PhnJ protein from numerous bacteria determined by maximum-likelihood analysis. Bootstrap values >50 are shown at the nodes where the trees from the three methods (maximum-likelihood, neighbour-joining and parsimony) were congruent; where the trees were incongruent, only bootstrap values from the likelihood analysis are shown. The scale bar represents amino acid substitutions per site. Numbers in parentheses are the IMG database gene object identifier, except for *Pseudomonas stutzeri*, which is the GenBank number. Full amino acid sequences are provided in the Supplementary Information.

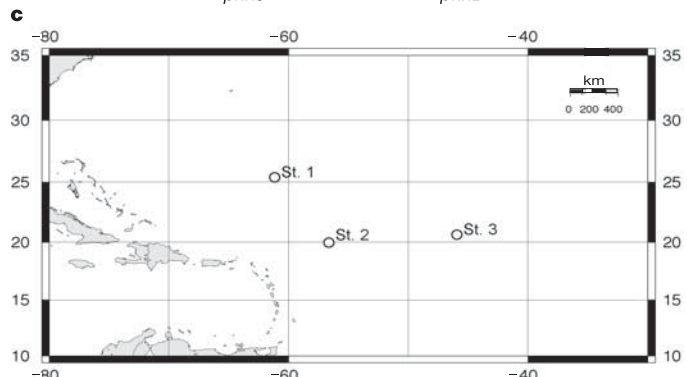
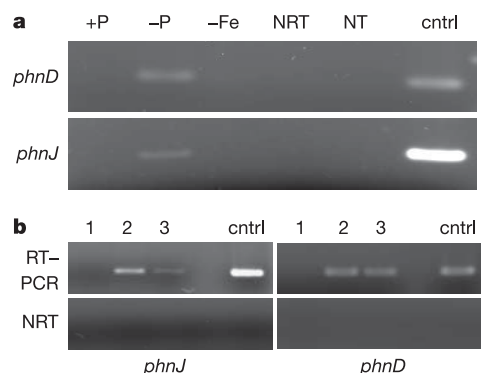


Figure 2 | Expression of *phnD* and *phnJ* in cultures and field populations of *T. erythraeum*. **a**, Expression of *phnD* (147 bp) and *phnJ* (280 bp) in *T. erythraeum* IMS101 from phosphate-replete (+P), phosphate-deficient (-P) and iron-deficient cultures (-Fe). All treatments were normalized to total RNA. Controls consisted of a no-RT sample (NRT), a no-template sample (NT) and genomic DNA from a phosphate-replete *T. erythraeum* IMS101 culture (cntrl). **b**, Expression of *phnD* (147 bp) and *phnJ* (280 bp) in extractions of field-collected *T. erythraeum* and in no-RT controls. Lane numbers indicate the station location. *T. erythraeum* IMS101 genomic DNA was used as a positive control (cntrl). All field samples were normalized to total RNA. **c**, Map of station locations in the western North Atlantic.

North Atlantic (Fig. 2). The field stations examined had low DIP concentrations (~ 6 nM) that did not significantly differ (Supplementary Table 2). Both genes were expressed and were detected at stations 2 and 3, but not at station 1 (Fig. 2). In no case was amplification observed in the no-RT controls, and expression of the photosynthetic gene *psbA* was detected at all stations (data not shown). The lack of *phnD* and *phnJ* expression at station 1 suggests that the *Trichodesmium* phosphorus physiology is different in this location. Although DIP is low at all stations, the lack of expression at station 1 could be explained by high rates of phosphorus cycling that would not be reflected in the DIP standing stock. Furthermore, gene expression reflects the nutritional history of a cell and its phosphorus quota, not necessarily its instantaneous nutrient environment. Together, these data show that *Trichodesmium* is both expressing the genes for phosphonate utilization in the field and regulating these genes *in situ*. To our knowledge, this is the first demonstration of differential expression of a functional gene in *Trichodesmium* field populations. The regulation and expression of genes involved in phosphonate transport and hydrolysis suggest that phosphonate compounds may be an important phosphorus source to *Trichodesmium* in this low DIP system.

DOP concentrations at all stations were substantially higher than DIP concentrations (Supplementary Table 2). Almost nothing is known about the concentration or distribution of specific dissolved phosphonate compounds in sea water²⁰, but there are many biological sources (such as phosphonolipids) and anthropogenic sources (such as pesticides) of distinct phosphonate compounds, suggesting that this pool is diverse^{17,20}. Considering the abundance of DOP and the likely diversity of the different phosphonate compounds within this pool, *Trichodesmium*, with its C–P lyase pathway, is well adapted for scavenging phosphorus from diverse phosphonates relative to other cyanobacteria.

With its dual roles of CO₂ and N₂ fixation and its prevalence in oligotrophic systems, *Trichodesmium* is an important model system for examining constraints on the marine cycling of carbon and nitrogen. One of these constraints is the bioavailability of phosphorus. Using data from the genome sequence, laboratory cultures and field samples, we have shown a capacity for phosphonate utilization in this important organism. The ability to utilize phosphonate compounds by the C–P lyase pathway would enable *Trichodesmium* to hydrolyse a diversity of phosphonate compounds. This broad substrate specificity is distinct from that mediated by phosphonate enzymes and thus represents an apparently unique niche adaptation that could explain the abundance of *Trichodesmium* in low phosphate systems such as the North Pacific Subtropical Gyre and the Sargasso Sea. Furthermore, these data indicate that phosphonates in the sea may be bioavailable to some marine diazotrophs: an observation that should be considered in further studies and models of marine biogeochemical cycling.

METHODS

Genomic observations. Genomic regions were visualized with the publicly available Integrated Microbial Genomes (IMG) interface on the Department of Energy Joint Genome Institute website (<http://img.jgi.doe.gov/pub/main.cgi>). The figure showing GC% was generated by using draft annotation sequences provided by F. Larimer (Oak Ridge National Laboratory) with the Artemis sequence viewing application (<http://www.sanger.ac.uk/Software/Artemis/>).

Phylogenetic analyses. Orthologues of *PhnJ* were obtained from the IMG interface, excluding the *P. stutzeri* HtxI protein, which was obtained from GenBank. Protein sequences were computer-aligned with ClustalW and hand-edited with the GCG Wisconsin package sequence editor (<http://www.accelrys.com/products/gcg/>). Neighbour-joining and parsimony phylogenies were generated with Paup 4.0b10 (<http://paup.csit.fsu.edu/>). The relative stability of topological elements was estimated with 100 bootstrap replicates for the maximum-likelihood analysis, and 1,000 replicates for the parsimony and neighbour-joining analyses. Gaps in the alignment and sites with consistency index values below 0.4 were excluded from the analyses. Maximum-likelihood phylogenies were generated with ProML and Consense applications in the Phylip

v.3.63 package using the Jones–Taylor–Thornton model of amino acid change. Neighbour-joining and parsimony analyses gave similar topologies.

Cultures. Axenic cultures of *T. erythraeum* IMS101 were grown in PMP medium (with or without added phosphate or iron) in a Sargasso Sea water base as described elsewhere²¹. Non-axenic cultures of *T. tenue*, *T. thiebautii* and *T. spiralis* were also maintained in PMP with the same Sargasso Sea water base. Growth was monitored using a Turner Designs fluorometer to assess relative pigment fluorescence as a proxy for biomass.

Field analyses. *Trichodesmium* colonies were collected with 130- μ m net tows within the top 10 m of the water column. Colonies of *T. erythraeum* were identified by morphology, picked quickly and washed several times in local seawater filtered through a 0.2- μ m sterile membrane. All colonies were stored frozen on filters in liquid nitrogen before analysis. Surface samples for inorganic phosphate were collected from either an acid-cleaned Teflon surface pump, or acid-cleaned GoFlo bottles at each station. These nutrient samples were filtered through 0.4- μ m membranes with an acid-cleaned manifold and stored frozen before analysis. Concentrations of DIP were assayed with the high-sensitivity MAGIC method²², and TDP was assayed by ultraviolet digestion²³ and a standard molybdate blue protocol²⁴. We calculated DOP as the difference between the TDP and DIP measurement.

Gene expression. RNA was extracted from *Trichodesmium* colonies on polycarbonate filters using a RiboPure Bacteria kit (Ambion) according to the manufacturer's instructions. The RNA was then treated with DNase I (Ambion) to remove trace amounts of genomic DNA. RNA concentrations were obtained with a NanoDrop ND-1000 spectrophotometer. All RNA preparations were checked for genomic contamination using PCR as described below and found to be negative. Total RNA was reverse-transcribed to single-stranded complementary DNA using an iScript cDNA Synthesis kit (Bio-Rad). To normalize the cDNA, the volume of RNA template added to the reaction mix from each field sample was adjusted to yield a final concentration of 0.95 ng μ l⁻¹. Additional reactions were done without reverse transcriptase (RT) to ensure the absence of genomic DNA in all cDNA preparations.

PCR amplification was done with 1 μ l of normalized cDNA template (0.0475 ng) in a final reaction mixture (15 μ l) containing 10 $\times$ PCR buffer (Bio-Rad), 0.25 μ M dNTPs (Bio-Rad), 50 pmol of each gene-specific 5' and 3' primer and 1 unit of Taq DNA polymerase (Bio-Rad). Optimal annealing temperatures were empirically determined to be 56.5 °C for both *phnD* and *phnJ*. The PCR conditions were 95 °C for 5 min; 35 cycles of 95 °C for 1 min, 56.5 °C for 1 min and 72 °C for 1 min; and 1 cycle of 72 °C for 10 min. Primer sets were designed to genes annotated in the *T. erythraeum* IMS101 genome. The *phnD* primer set was designed to amplify a 147-bp region (5' end, CAGACTCCCGTTCGGAATTA; 3' end, CACCGGCATAGTCAGTAGCA). The *phnJ* primer set was designed to amplify a 280-bp region (5' end, AAGACCA-GAGCCCCACCTAT; 3', end GATACCCCGGACAAGTTT). PCR products were resolved on a 2% agarose gel, stained with ethidium bromide, and imaged with a Gel Logic 440 Imaging System (Kodak).

DNA sequencing. DNA was extracted by using Instagene Matrix (Bio-Rad) according to the manufacturer's instructions and amplified as described above. Gene products were excised from agarose gels and purified with a QIAquick gel extraction kit (Qiagen). DNA sequencing was done using an ABI 3730xl sequencer at the Josephine Bay Paul Center of the Marine Biological Laboratory according to the facility's protocols. In brief, 5–20 ng of template was combined with 7.5 pmol of primer (*phnJ* or *phnD*), and 1 μ l of ABI BigDye Terminator v3.1 cycle sequencing mix in a final volume of 6 μ l. The cycle sequencing program consisted of 25 cycles of 96 °C for 10 s, 50 °C for 5 s, and 60 °C for 4 min. Unincorporated labelled nucleotides were removed by isopropanol precipitation, and the reaction products were resuspended in formamide for analysis. We edited the sequences with Sequencer v4.2.2 and aligned them with Macvector v8.0.

Received 11 July; accepted 6 September 2005.

1. Wu, J., Sunda, W. G., Boyle, E. A. & Karl, D. M. Phosphate depletion in the Western North Atlantic Ocean. *Science* **289**, 759–762 (2000).
2. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. G. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
3. Montoya, J. P. et al. High rates of N₂ fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**, 1027–1031 (2004).
4. Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. & Carpenter, E. J. *Trichodesmium*, a globally significant marine cyanobacterium. *Science* **276**, 1221–1229 (1997).
5. Carpenter, E. J. & Romans, K. M. Major role of the cyanobacterium *Trichodesmium* in nutrient cycling in the North Atlantic Ocean. *Science* **254**, 1356–1358 (1991).

6. Sañudo-Wilhelmy, S. A. *et al.* Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. *Nature* **411**, 66–69 (2001).
7. Sañudo-Wilhelmy, S. A. *et al.* The impact of surface-adsorbed phosphorus on phytoplankton Redfield stoichiometry. *Nature* **432**, 897–901 (2004).
8. Ammerman, J. W., Hood, R. R., Case, D. A. & Cotner, J. B. Phosphorus deficiency in the Atlantic: an emerging paradigm in oceanography. *Eos* **84**, 165–170 (2003).
9. Clark, L. L., Ingall, E. D. & Benner, R. Marine phosphorus is selectively remineralized. *Nature* **393**, 426 (1998).
10. Kolowitz, L. C., Ingall, E. D. & Benner, R. Composition and cycling of marine organic phosphorus. *Limnol. Oceanogr.* **46**, 309–320 (2001).
11. Stihl, A., Sommer, U. & Post, A. F. Alkaline phosphatase activities among populations of the colony-forming, diazotrophic cyanobacterium *Trichodesmium* spp. (Cyanobacteria) in the Red Sea. *J. Phycol.* **62**, 310–317 (2001).
12. Dyhrman, S. T., Webb, E., Anderson, D. M., Moffett, J. & Waterbury, J. Cell specific detection of phosphorus stress in *Trichodesmium* from the Western North Atlantic. *Limnol. Oceanogr.* **47**, 1823–1836 (2002).
13. Paytan, A., Cade-Menun, B. J., McLaughlin, K. & Faul, K. L. Selective phosphorus regeneration of sinking marine particles: evidence from ³¹P-NMR. *Mar. Chem.* **82**, 55–70 (2003).
14. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424**, 1037–1042 (2003).
15. Moore, L. R., Ostrowski, M., Scanlan, D. J., Feren, K. & Sweetsir, T. Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. *Aquat. Microb. Ecol.* **39**, 257–269 (2005).
16. Wanner, B. L. Molecular genetics of carbon–phosphorus bond cleavage in bacteria. *Biodegradation* **5**, 175–184 (1994).
17. Kononova, S. V. & Nesmeyanova, M. A. Phosphonates and their degradation by microorganisms. *Biochemistry (Moscow)* **67**, 220–233 (2002).
18. White, A. W. & Metcalf, W. M. Two C–P lyase operons in *Pseudomonas stutzeri* and their roles in the oxidation of phosphonates, phosphite and hypophosphite. *J. Bact.* **186**, 4730–4739 (2004).
19. Lundgren, P., Janon, S., Jonasson, S., Singer, A. & Bergman, B. Unveiling of novel radiations within *Trichodesmium* cluster by *hetR* gene sequence analysis. *Appl. Environ. Microbiol.* **71**, 190–196 (2005).
20. Karl, D. M. & Björkman, K. M. in *Biogeochemistry of Marine Dissolved Organic Matter* (eds Hansell, D. & Carlson, C.) 249–366 (Elsevier Science, Boston, USA, 2002).
21. Webb, E. A., Moffett, J. W. & Waterbury, J. B. Iron stress in open-ocean cyanobacteria (*Synechococcus*, *Trichodesmium*, and *Crocospheera* spp.): identification of the IdiA protein. *Appl. Environ. Microbiol.* **67**, 5444–5452 (2001).
22. Karl, D. M. & Tien, G. MAGIC: A sensitive and precise method for measuring dissolved phosphorus in aquatic environments. *Limnol. Oceanogr.* **37**, 105–116 (1992).
23. Armstrong, F. A. J., Williams, P. M. & Strickland, J. D. H. Photo-oxidation of organic matter in sea water by ultra-violet radiation, analytical and other applications. *Nature* **211**, 481–483 (1966).
24. Koroleff, F. in *Methods of Seawater Analysis* (eds Grashoff, K., Ehrhardt, M. & Kremling, K.) 125–138 (Verlag Chemie, Weinheim, Denmark, 1983).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the captain and crew of the RV *Oceanus*, I. Ehrenreich and R. Wisniewski for assistance; and V. Edgecomb and F. Valois for suggestions on the manuscript. The *T. erythraeum* IMS101 genome was produced by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov/>). This work was supported by the National Science Foundation Biological Oceanography Program, the Woods Hole Oceanographic Institution and the Princeton Center for Biomineral Chemistry.

Author Contributions S.T.D. and E.A.W. designed the research and wrote the paper; S.T.D., J.W.M. and E.A.W. performed the field work; P.D.C., S.T.D., S.T.H., E.D.O., J.B.W. and E.A.W. performed the laboratory work; and E.A.W. performed the genome searches and phylogenetic analyses.

Author Information The sequences of *phnJ* and *phnD* from the different *Trichodesmium* species have been deposited in GenBank with accession numbers DQ176437–DQ176442. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E. A. W. (ewebb@whoi.edu).

LETTERS

Computer optimization of a minimal biped model discovers walking and running

Manoj Srinivasan¹ & Andy Ruina¹

Although people's legs are capable of a broad range of muscle-use and gait patterns, they generally prefer just two. They walk, swinging their body over a relatively straight leg with each step, or run, bouncing up off a bent leg between aerial phases. Walking feels easiest when going slowly, and running feels easiest when going faster. More unusual gaits seem more tiring. Perhaps this is because walking and running use the least energy^{1–7}. Addressing this classic¹ conjecture with experiments^{2,3} requires comparing walking and running with many other strange and unpractised gaits. As an alternative, a basic understanding of gait choice might be obtained by calculating energy cost by using mechanics-based models. Here we use a minimal model that can describe walking and running as well as an infinite variety of other gaits. We use computer optimization to find which gaits are indeed energetically optimal for this model. At low speeds the optimization discovers the classic inverted-pendulum walk^{8–13}, at high speeds it discovers a bouncing run^{12,13}, even without springs, and at intermediate

speeds it finds a new pendular-running gait that includes walking and running as extreme cases.

One way of characterizing gaits is by the motions of the body (Fig. 1a). In these terms, walking seems well caricatured¹³ (Fig. 1b) by the hip joint going from one circular arc to the next with push-off and heel-strike impulses in between. Similarly, running could be caricatured by a sequence of parabolic free-flight arcs (Fig. 1c), with impulses from the ground at each bounce^{14–17}.

Why do people not walk or even run with a smooth level gait⁸, like a waiter holding two cups brim-full of boiling coffee? Why do people select walking and running from the other possibilities? We address such questions by modelling a person as a machine describable with the equations of newtonian mechanics. The basic approximations are: first, that humans have compact bodies and light legs; second, that gait choice is based on energy optimization^{1,4}; and third, that energy cost is proportional to muscle work^{2,4,8}. We use a simplification of previous models^{4,6,7}, perhaps the simplest mechanical model

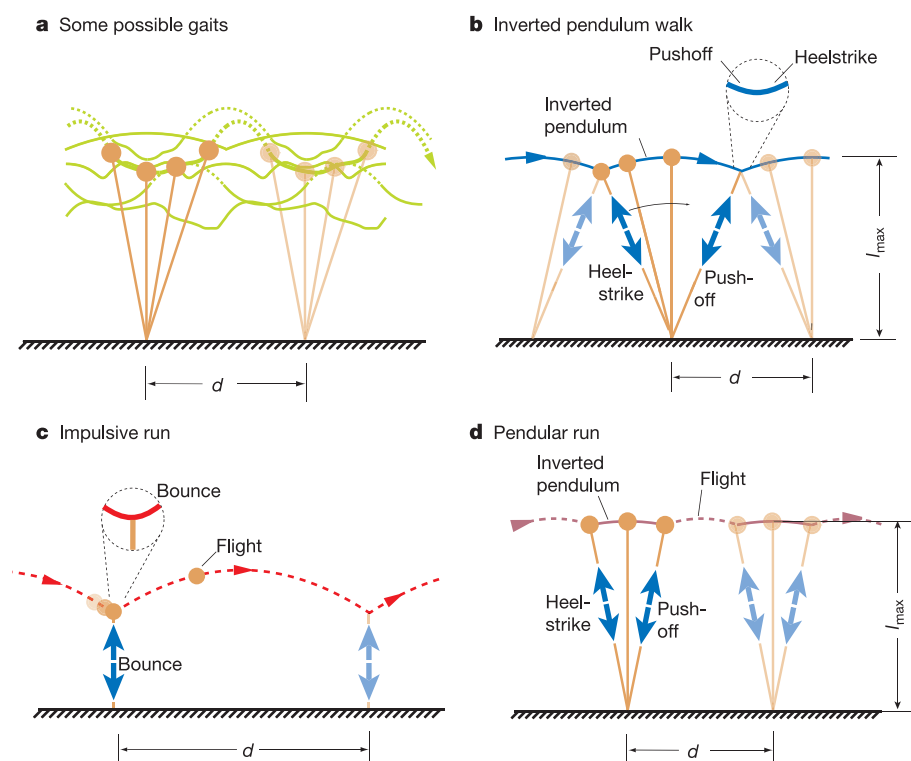


Figure 1 | Body motion in human gaits. **a**, Trajectories of the centre of mass for a few possible gaits. Solid lines, stance; dotted lines, flight. **b**, Trajectory for inverted-pendulum walking. **c**, Trajectory for impulsive running.

d, Trajectory for a new gait: pendular running. At least one of the gaits **b**, **c** and **d** turns out to use less work than any other candidates (for example, from **a**), according to the calculations here.

¹Bio-robotics and Locomotion Laboratory, Theoretical and Applied Mechanics, Cornell University, 306 Kimball Hall, Ithaca, New York 14853, USA.

that is capable of exhibiting a broad range of gaits that includes walking and running. Although the model is a mechanical abstraction that is not physically realizable, it is subject to the laws of physics. Because of its simplicity, the model is amenable to interpretation. It can also be studied with exhaustive and accurate simulation experiments, far beyond what is possible with human subjects.

We wish to find how a person can get from one place to another with the least muscle work W (Methods). We treat the body as a point mass m at position (x, y) at time t (Fig. 2a). The legs are massless and therefore, when not in ground contact, they can be oriented, lengthened and shortened with no energy cost. The fluctuations of the leg length $l(t)$ due to flexion of the hip, knee and ankle are incorporated in a single telescoping axial actuator⁴ that carries a compressive time-varying force $F = F(t)$. For simplicity, we seek an explanation of gait choice with no essential dependence on elastic energy storage; we assume no springs (tendons) in series or parallel with the actuators.

We assume that during the stance phase, when a foot is in contact with the rigid level ground, that it does not slip. At most one foot can be in contact with the ground at a time. During stance, both gravity mg and F act on the body (Fig. 2a). During the flight phase, when neither leg touches the ground, gravity is the only force. We seek periodic motions, in which each step is like the previous step. The left and right legs have identical force and length profiles. A single step consists of one stance phase (possibly short, as in high-speed running) and one flight phase (possibly of zero duration, as in walking).

A gait is characterized by the position and velocity of the body at the start of a stance phase relative to the stance foot, by the step period, and by $F(t)$. Given these, we can integrate the newtonian equations of motion forwards in time to find the body trajectory and leg length as functions of time (including the maximum leg length $l_{\max}$). At the end of the step, we assume that the next foot is placed on

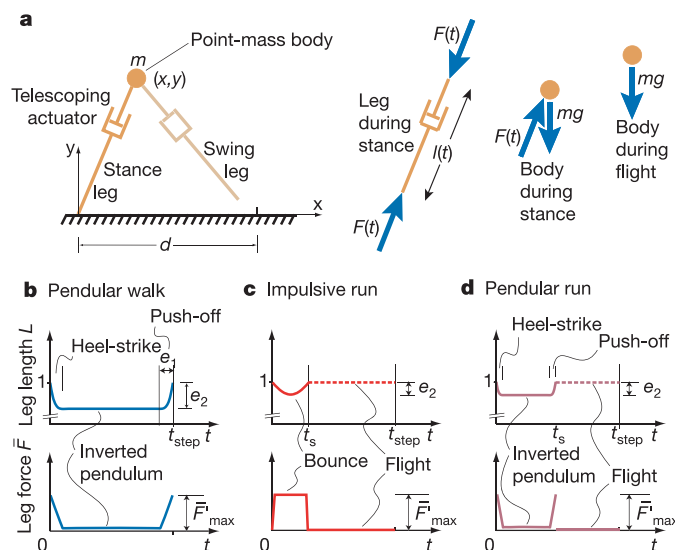


Figure 2 | Point-mass biped model and its optimal solutions. **a**, The configuration shown is part way through the stance phase. The next stance leg is oriented to prepare for a new contact at a distance d from the last. **b–d**, Dimensionless force and length shown as functions of dimensionless time, for the three optimal gaits (**b**, pendular walk; **c**, impulsive run; **d**, pendular run), before full convergence of the numerical optimization. The finite forces in the figures are approximations to the converged impulsive (collisional) forces. In the extrapolated optimum, as the grid size $h \rightarrow 0$ and the allowed force upper bound $F_{\max} \rightarrow \infty$, the optimizations find that $e_1, e_2 \rightarrow 0$ and that the maximum forces used go to infinity (Methods). In these limits the walking gait (**b**) is an inverted pendulum with heel-strike and push-off impulses, the running gait (**c**) is an impulsive bounce between free flights, and the pendular run (**d**) has constant-length pendulum phases and flight phases separated by impulses.

the ground at the same position relative to the body as at the start. We can thus calculate the step length d , the average forward speed v , and the work done by the leg per unit weight and distance $C = W/(mgd)$. For random $F(t)$, the final body height and velocity generally do not match the starting conditions and therefore do not generate a periodic gait. Nonetheless, by appropriately varying $F(t)$ we can find infinitely many periodic gaits (Fig. 1a) with all manner of complicated trajectories (Methods). Of those periodic gaits, we wish to find those that minimize the cost C .

The optimal solutions have cost arbitrarily close to zero unless the optimization is further constrained. The cost can be made arbitrarily small by growing the leg length (and the locomotion becomes akin to the rolling of a giant multi-spoked wheel), so we set the maximum length to be $l_{\max}$, representing the leg length. Because we have no leg-swing cost, C can be reduced to zero by taking very small steps^{6,12,18}, so we optimize for various fixed values of step length d . Finally, C has a non-anthropomorphic lower bound (corresponding to standing on one leg for an infinite time mid-step), approached as the average speed v goes to zero, so we constrain v .

After non-dimensionalizing using m , g and $l_{\max}$, no free parameters remain. We seek solutions as two conditions are varied: the dimensionless average speed $V = v/\sqrt{gl_{\max}}$ (V^2 is the so-called Froude number) and the dimensionless step length $D = d/l_{\max}$. For given values of V and D , the optimal periodic gait is determined with numerical optimal control methods that are more or less standard (Methods).

All optimizations converged towards one of three stereotypical collisional gaits, depending on V and D , but never to a smooth collisionless gait. First, at low V , the classic inverted-pendulum walking gait (Figs 1b and 2b) is optimal. Second, at high V , an impulsive running gait is optimal (Figs 1c and 2c). Third, at intermediate V , a new gait, pendular running (Figs 1d and 2d), is optimal. Pendular running has a flight phase between extended inverted-pendulum stance phases. Pendular running is a generalization of, and a connection between, walking and running: with no flight phase it is inverted-pendulum walking; with an infinitesimal pendular phase it is impulsive running.

The numerical optimization, unbiased by an expectation of what the optimal gaits might be, has thus discovered the classic gaits that caricature walking and running. The new third gait might be the model's way of running with a non-zero stance phase, given the

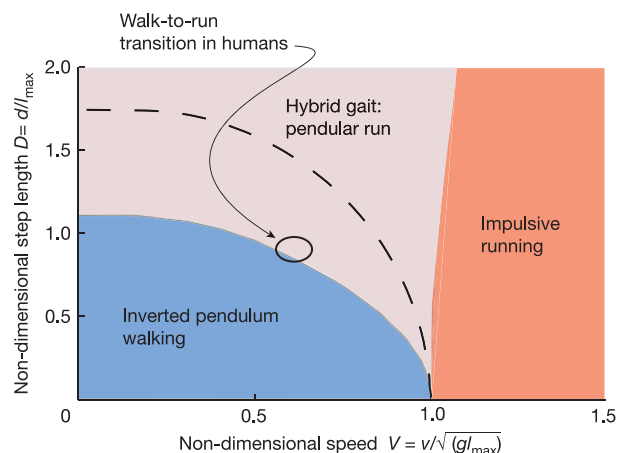


Figure 3 | The regions in which each of the three collisional gaits are optimal. Inverted-pendulum walking ceases to be locally optimal at the pendular-run interface. The oval indicates the approximate speed and step length range at which humans switch from walking to running^{19,20}. The dashed line indicates where compression-only inverted-pendulum walking becomes mechanically infeasible (typically approximated⁸ as $V = 1$, which is correct for small D). At the right part of the intermediate region, the pendular run is almost impulsive running; at the left edge, it is almost inverted-pendulum walking.

model's lack of tendons. A tentative prediction would be the existence of a ground force versus time curve with two humps during the stance phase for, perhaps, weak or obese people running slowly. The respective regions of optimality of the three gaits are shown in Fig. 3.

Alexander^{8,10} argued that inverted-pendulum walking is limited to those speeds at which the centripetal acceleration of a body pivoting over a straight leg is less than gravity, ensuring that the body does not vault off the ground. However, walking becomes energetically non-optimal at speeds lower than the above limit^{8,10} (Fig. 3). Indeed, people switch from a walk to run^{19,20} at about $V = 0.65$ and $D = 0.95$, close to the boundary at which walking ceases to be optimal (Fig. 3) in this model.

The numerical optimization results are buttressed by heuristic considerations. The cost C is an integral of the leg power (Methods). There are two ways of setting this power to zero: setting $\dot{l} = 0$ (corresponding to inverted-pendulum motion) or setting $F = 0$ (corresponding to free flight). Thus, the flight phase ($F = 0$) of running is an energy-saving analogue of the pendular ($\dot{l} = 0$) motion of walking; both phases involve no work. All the work is crowded into brief impulses at appropriate times.

Inverted-pendulum walking, pendular running and impulsive

running all have work-free motions, punctuated by impulses (collisions). The costs of these collisional gaits can be calculated directly^{10–12}. For inverted-pendulum walking, positive work performed during push-off is evaluated as the difference in kinetic energy just before and after the push-off^{8,11,12}. $C_{\text{walking}} = DV_1^2/(8 - 2D^2)$, where V_1 is the magnitude of the velocity vector just before push-off. For impulsive running, cost is equal to the vertical kinetic energy that is lost and regained in every bounce^{12,13} ($C_{\text{running}} = D/8V^2$). For a given V and small values of D , the cost for the collisional gaits is proportional¹² to the square of the kink-angle in the trajectory (Fig. 4c). The energetic trade-off between inverted-pendulum walking and impulsive running (Fig. 4a, b) can be understood as a minimization of collision angles¹² for a specific step length D . At low speeds the circular arc of walking has shallower collisions than the parabolic free-flight of running, and at high speeds the situation is reversed (Fig. 4c).

The optimizations here show that smooth collisionless gaits require more work than the optimal collisional gaits. For example, consider a flat walk^{8,10}, in which the body moves at constant height. This gait has^{8,10} $C_{\text{flat}} = D/8\sqrt{1 - D^2}$. Figure 4a, b shows that the exceptionally smooth, flat walk is never optimal (Methods). Recent human experiments^{21,22} also show that a flat walk uses more energy than normal walking.

As has been found for a gait model that assumes collisions *a priori*¹², the more general model here shows that it is advantageous to simulate elasticity during running, even with no genuine elasticity (tendons). Indeed, real human legs do approximately simulate an elastic spring during running^{16,17}. More generally, the model here, as well as simpler models^{4,8,12}, indicates that the energetic utility of running probably does not depend on genuine elasticity in the legs. However, such elasticity, neglected here, would further decrease the cost of running^{4,6,9}, supporting the idea²³ that human ancestors could have started to run before the modern human long Achilles tendon was fully evolved.

To maximize simplicity of calculation and interpretation, we have neglected various crucial features including a cost for leg-swing^{12,18,24}, a more realistic model of muscle cost^{7,25}, allowance of a non-infinitesimal double-stance phase^{4,6,7}, elastic and dissipative elements in series with the actuator^{4,6,7,23}, the possibility of higher-period gaits (for example skipping²⁶), an extended foot instead of a point foot²⁷, and other anatomical realism²⁷.

The simplest way of including a leg-swing cost would be to assume that it is a function of frequency and amplitude which is independent of gait. The leg-swing cost is then a function of V and D , has no effect on which gait uses less energy at a given V and D , and therefore has no effect on which gait is optimal at that V and D . Figure 3 would be exactly unchanged. The simplest way of incorporating elastic recovery is to assume that a fixed fraction of the leg work is from elastic energy storage and hence should have no cost in the optimization. This would scale the costs of all gaits by the same constant (less than 1) and would therefore have no effect on any of the relative costs of various gaits. Thus, leg-swing and elastic-recovery effects can affect gait choice only through more complex dependences.

We do not know which neglected effects are the most important for explaining the deviations of observed human behaviour from the model predictions here, particularly the prediction of the pendular-running gait, which seems little used by humans. Nonetheless, this model, having no free parameters, might most simply explain why we choose walking and running over the plethora of other possible gaits.

METHODS

Formulation. The governing equations are

$$m\ddot{x} = F(x - x_c)/l, m\ddot{y} = -mg + Fy/l \quad (1)$$

for stance with duration t_s , and

$$\ddot{x} = 0, \ddot{y} = -g \quad (2)$$

for flight with duration t_f , where $l = \sqrt{(x - x_c)^2 + y^2}$. Time $t = 0$ is the

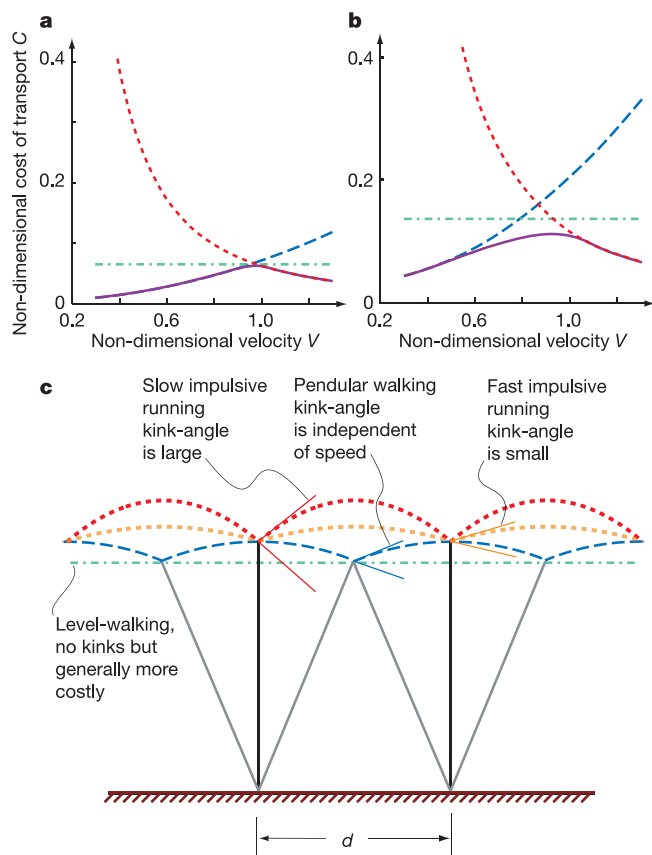


Figure 4 | Cost of transport versus speed. **a**, For small D ($= 0.50$), all periodic gaits (that do not involve leg tension) have nearly equal costs near $V = 1$. Inverted-pendulum walking is optimal at low speeds, pendular running at a narrow range of intermediate speeds, impulsive running at high speeds, and flat walking is never optimal. **b**, However, for large D ($= 1.00$) and for $V \approx 0.8$ – 0.9 , flat walking, perhaps like a ‘Groucho walk’³⁰, although not optimal, has lower cost than both inverted-pendulum walking and impulsive running. The colours used in **a** and **b** indicate the following gaits: red, impulsive running; blue, pendular walking; green, level walking; purple, optimal gait. **c**, Body trajectories for a pendular walking gait (blue; kink angle is independent of speed), a low-speed impulsive running gait (red; kink angle is large), a high-speed impulsive running gait (orange; kink angle is small) and level walking (green; no kinks, but generally more costly), all with the same step length.

beginning of a stance phase with foot-contact point $x_c = 0$. The initial conditions are $x(0) = x_0$, $y(0) = y_0$, $\dot{x}(0) = \dot{x}_0$ and $\dot{y}(0) = \dot{y}_0$. At $t = t_f + t_s$, periodicity requires that $x_f = x_0 + d$, $y_f = y_0$, $\dot{x}_f = \dot{x}_0$ and $\dot{y}_f = \dot{y}_0$. The numerical integration then determines v , d , $l_{\max}$ and C . For given $l_{\max}$, d and v , we seek the control strategy $(x_0, \dot{x}_0, y_0, \dot{y}_0, F(t), t_s)$ that minimizes the work-based specific mechanical cost of transport

$$C = \int_0^{t_{\text{step}}} [F(t)\dot{l}]^+ dt / mgd \quad (3)$$

where $[\cdot]^+$ is non-zero only for positive values ($[p]^+ = p$ if $p \geq 0$ and $[p]^+ = 0$ if $p < 0$). The only cost is for mechanical work ($dW = Fdl$).

Numerical solution of the optimal control problem. We non-dimensionalize all quantities by $l_{\max}$, M and g . We seek $(X_0, \dot{X}_0, Y_0, \dot{Y}_0, \bar{F}(\tau), \tau_s) = (x_0/l_{\max}, \dot{x}_0/\sqrt{gl_{\max}}, y_0/l_{\max}, \dot{y}_0/\sqrt{gl_{\max}}, F(t)/mg, t_s\sqrt{g/l_{\max}})$, where τ is the non-dimensional time, that produce the optimal periodic gait with given V and D , and with the non-dimensional step-length satisfying $0 \leq L(\tau) \leq 1$.

The infinite-dimensional search space for this optimization problem contains the set of all possible functions $\bar{F}(\tau)$. We restrict our search to the set of piecewise linear functions, defined on an evenly spaced time-grid ($0 = \tau_0, \tau_1, \tau_2, \dots, \tau_N = \tau_s$), with grid spacing $\tau_i - \tau_{i-1} = h = \tau_s/N$. So the search space becomes $z = (X_0, \dot{X}_0, Y_0, \dot{Y}_0, \bar{F}_{i=0,\dots,N}, \tau_s)$, where $\bar{F}_i(\tau) = \bar{F}(\tau_i)$. The linear constraints are $\varepsilon \leq \tau_s \leq \tau_{\text{step}}$, $\bar{F}_{\min} \leq \bar{F}_i \leq \bar{F}_{\max}$. We need $\varepsilon > 0$ because a periodic step requires a stance phase. In addition, although the forces are allowed to be unbounded conceptually, for numerical optimization they need to be bounded: we choose a bound $\bar{F}_{\max} \gg 1$ and $\bar{F}_{\min} = 0$. Ultimately $\bar{F}_{\max}$ is allowed to grow arbitrarily, so that it is not a parameter in the solutions we present. Interestingly, choosing $\bar{F}_{\min} < 0$, allowing tensional leg-forces, does not affect the optima. The leg-length constraint, $0 \leq L(\tau) \leq 1$, is enforced at the grid points $\tau = \tau_i$. Gait periodicity is another nonlinear constraint.

For given z , C and the constraint violations are evaluated by integration of the differential equations. $C(z)$ is to be minimized subject to the various linear and nonlinear equality and inequality constraints: $g_{\text{eq}}(z) = 0$ and $g_{\text{ineq}}(z) \leq 0$. We smooth $C(z)$ with h as a smoothing parameter. We used a particularly robust implementation of Sequential Quadratic Programming (SQP)²⁸ for the optimization.

Convergence to the idealized collisional gaits is discovered by letting $N \rightarrow$ large, $\bar{F}_{\max} \rightarrow$ large and $\varepsilon \rightarrow$ small. At high V , if $\bar{F}_{\max}$ is set large enough for a given ε , $\bar{F}_{\max}$ has no effect on C . The optimization then always finds $\tau_s = \varepsilon$ as $\varepsilon \rightarrow 0$, thus converging to impulsive running. We assure ourselves of the convergence to the collisional walking by Richardson extrapolation. That is, we solve the problem for grids of sizes $N = N_1, N_2, N_3, \dots$, assuming that the cost is a smooth function of N^{-1} , and extrapolating the cost to $N^{-1} \rightarrow 0$. $\bar{F}_{\max}$ is maintained high enough and ε low enough to be unused constraints. The ODE solutions are accurate to about 10^{-14} over a grid interval (obtained by integrating from grid-point to grid-point with an adaptive RK-45 method, benchmarked by a Taylor-series method) and accurate to less than $10^{-14}N$ over the whole step. We thus avoid significant sources of error not related to the finiteness of N and can therefore treat the convergence as dependent only on N . The convergence is observed to be linear in N^{-1} . The linearly extrapolated limit of the sequence of C values is found to differ from the cost of the corresponding analytically determined inverted-pendulum collisional walking gait by a relative error of about 10^{-3} .

For each V and D , multiple optimization runs, each started with a different initial seed, all converged towards the same control strategy, indicating the likely uniqueness and globality of each collisional minimum. To determine the regions in which each gait is optimal more precisely (Fig. 3) we repeated the optimization over the space of (analytically calculable) collisional gaits.

Pontryagin's maximum principle. Pontryagin's maximum principle²⁹ can be used over the stance phase, neglecting the leg-length constraint, to get necessary conditions on the optimal solutions. This calculation shows that during stance, if the optimal control is not singular, the leg-forces must be maximum ($F_{\max}$, apparently corresponding to heel-strike or push-off), or zero (stance simulating flight by having no force). This much agrees with our full optimizations and heuristics. The pendular stance portions we found, with $\dot{l} = 0$, seem to be singular arcs of the optimal control.

Received 29 June; accepted 3 August 2005.

Published online 11 September 2005.

1. Borelli, G. A. *De Motu Animalium, Pars Prima (On the Movement of Animals)* (Angeli Bernabo, Rome, 1680).

2. Margaria, R. *Biomechanics and Energetics of Muscular Exercise* (Clarendon, Oxford, 1977).
3. Hoyt, D. F. & Taylor, C. R. Gait and the energetics of locomotion in horses. *Nature* **292**, 239–240 (1981).
4. Alexander, R. McN. Optimum walking techniques for quadrupeds and bipeds. *J. Zool. Lond.* **192**, 97–117 (1980).
5. Alexander, R. McN. Optimization and gaits in the locomotion of vertebrates. *Physiol. Rev.* **69**, 1199–1227 (1989).
6. Alexander, R. McN. A model of bipedal locomotion on compliant legs. *Phil. Trans. R. Soc. Lond. B* **338**, 189–198 (1992).
7. Minetti, A. E. & Alexander, R. McN. A theory of metabolic costs for bipedal gaits. *J. Theor. Biol.* **186**, 467–476 (1997).
8. Alexander, R. McN. in *Perspectives in Experimental Biology* Vol. 1 (ed. Davies, P. S.) 493–504 (Pergamon, New York, 1976).
9. Cavagna, G. A., Heglund, N. C. & Taylor, C. R. Mechanical work in terrestrial locomotion: two basic mechanisms for minimizing energy expenditure. *Am. J. Physiol.* **233**, R243–R261 (1977).
10. Alexander, R. McN. Simple models of human motion. *Appl. Mech. Rev.* **48**, 461–469 (2003).
11. Kuo, A. D. Energetics of actively powered locomotion using the simplest walking model. *J. Biomech. Eng.* **124**, 113–120 (2002).
12. Ruina, A., Bertram, J. E. A. & Srinivasan, M. A collisional model of the energetic cost of support work qualitatively explains leg sequencing in walking and galloping, pseudo-elastic leg behavior in running and the walk-to-run transition. *J. Theor. Biol.* [online], 14 June 2005 (doi:10.1016/j.jtbi.2005.04.004).
13. Kuo, A. D., Donelan, J. M. & Ruina, A. Energetic consequences of walking like an inverted pendulum: step-to-step transitions. *Exercise Sport Sci. Rev.* **3**, 88–97 (2005).
14. Rashevsky, N. Studies in the physico-mathematical theory of organic form. *Bull. Math. Biophys.* **6**, 1–59 (1944).
15. Alexander, R. McN. *Elastic Mechanisms in Animal Movement* (Cambridge Univ. Press, Cambridge, 1988).
16. McMahon, T. A. & Cheng, G. C. The mechanics of running: how does stiffness couple with speed? *J. Biomech.* **23** (Suppl. 1), 65–78 (1990).
17. Blickhan, R. & Full, R. J. Similarity in multilegged locomotion: bouncing like a monopode. *J. Comp. Physiol. A* **173**, 509–517 (1993).
18. Kuo, A. D. A simple model predicts the step length-speed relationship in human walking. *J. Biomech. Eng.* **123**, 264–269 (2001).
19. Thorstensson, A. & Robertson, H. Adaptations to changing speed in human locomotion: speed of transition between walking and running. *Acta Physiol. Scand.* **131**, 211–214 (1987).
20. Minetti, A. E., Ardigo, L. P. & Saibene, F. The transition between walking and running in humans: metabolic and mechanical aspects at different gradients. *Acta Physiol. Scand.* **150**, 315–323 (1994).
21. Farley, C. T. & Ortega, J. D. Minimizing center of mass vertical movement increases metabolic cost in walking. *J. Appl. Physiol.* [online], 28 July 2005 (doi:10.1152/jappphysiol.00103.2005).
22. Gordon, K., Ferris, D. & Kuo, A. *Proc. 27th Annual Meeting, American Society of Biomechanics* Abstr. 113 (American Society of Biomechanics, Toledo, Ohio, 2003).
23. Bramble, D. M. & Lieberman, D. E. Endurance running and the evolution of *Homo*. *Nature* **432**, 345–352 (2004).
24. Marsh, R. L. *et al.* Partitioning the energetics of walking and running: swinging the legs is expensive. *Science* **303**, 80–83 (2004).
25. Kram, R. & Taylor, C. R. Energetics of running: a new perspective. *Nature* **346**, 265–267 (1990).
26. Minetti, A. E. The biomechanics of skipping gaits: a third locomotor paradigm? *Proc. R. Soc. Lond. B* **265**, 1227–1235 (1998).
27. Anderson, F. C. & Pandy, M. G. Dynamic optimization of human walking. *J. Biomech. Engng* **123**, 381–390 (2001).
28. Gill, P. E., Murray, W. & Saunders, M. A. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM J. Optim.* **12**, 979–1006 (2002).
29. Bryson, A. E. & Ho, Y. C. *Applied Optimal Control* (John Wiley, New York, 1975).
30. Bertram, J. E. A., D'Antonio, P., Pardo, J. & Lee, D. V. Pace-length effects in human walking: 'Groucho gaits' revisited. *J. Mot. Behav.* **34**, 309–318 (2002).

Acknowledgements We thank J. Bertram for extensive discussions on related topics; J. Burns, A. Chatterjee, P. Holmes, A. Schwab, S. Strogatz, S. van Nieuwenhuis and S. Walcott for editorial and technical comments.

Author Contributions M.S. formulated the mechanical model and performed the numerical optimizations. This was partly informed by extensive discussions with A.R. during the writing of ref. 12, and from a course taught by A.R. The authors contributed equally to the analytic calculations, the interpretation of the results and to the writing of the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.S. (ms285@cornell.edu).

LETTERS

Complex social behaviour derived from maternal reproductive traits

Gro V. Amdam^{1,2}, Angela Csondes³, M. Kim Fondrk¹ & Robert E. Page Jr¹

A fundamental goal of sociobiology is to explain how complex social behaviour evolves¹, especially in social insects, the exemplars of social living. Although still the subject of much controversy², recent theoretical explanations have focused on the evolutionary origins of worker behaviour (assistance from daughters that remain in the nest and help their mother to reproduce) through expression of maternal care behaviour towards siblings^{3,4}. A key prediction of this evolutionary model is that traits involved in maternal care have been co-opted through heterochronous expression of maternal genes⁵ to result in sib-care, the hallmark of highly evolved social life in insects⁶. A coupling of maternal behaviour to reproductive status evolved in solitary insects, and was a ready substrate for the evolution of worker-containing societies^{3,4,7,8}. Here we show that division of foraging labour among worker honey bees (*Apis mellifera*) is linked to the reproductive status of facultatively sterile females. We thereby identify the evolutionary origin of a widely expressed social-insect behavioural syndrome^{1,5,7,9}, and provide a direct demonstration of how variation in maternal reproductive traits gives rise to complex social behaviour in non-reproductive helpers.

Worker honey bees change the tasks that they perform with age¹⁰. This behaviour results in a division of labour that is age-associated¹¹. Workers usually make a transition from working in the nest to foraging in their second or third week of life¹², and foragers often specialize in collecting nectar or pollen. Recent studies have identified a suite of traits that differ between nectar and pollen foragers⁹. These traits are affected by a pleiotropic genetic network¹³, and it has been suggested that this pleiotropy can be explained if a reproductive regulatory network was co-opted by natural selection to differentiate the foraging behaviour of the facultatively sterile workers⁷. This hypothesis emerged from studies of honey bees that were selected to collect and store high (the high-hoarding strain) or low (the low-hoarding strain) amounts of pollen¹⁴. Traits of the strains diverge, so that high pollen-hoarding bees switch from nest tasks to foraging earlier in life, and are more likely to collect pollen and carry larger pollen loads. Bees from the high pollen-hoarding strain are more likely than bees from the low pollen-hoarding strain to collect water and nectar with low sugar concentration, and at emergence they have higher haemolymph (blood) levels of juvenile hormone and vitellogenin protein⁷. Pollen foraging is a maternal reproductive behaviour in solitary bees, and non-reproductive females feed mainly on nectar¹⁵. Elevated juvenile hormone levels cause physiological and behavioural changes during the reproductive maturation of many insects^{7,16,17}, and vitellogenin is a conserved yolk precursor synthesized by most oviparous females¹⁸. Therefore, the evidence from pollen-hoarding strains suggests that nectar-foraging bees display a non-reproductive phenotype, whereas pollen foragers display the ancestral maternal character state of solitary species⁷. As a

consequence, the foraging division of labour between worker bees would be derived from variation in maternal reproductive traits. Validation of this hypothesis, however, requires the demonstration of a relationship between the reproductive status and the foraging behaviour of honey bee workers⁷.

We addressed the debate on the origin of complex social behaviour by first inspecting the number of ovarioles (egg-forming filaments in the ovary) in newly emerged workers from the previously examined⁷ high and low pollen-hoarding strains. Developmental differentiation of ovariole number¹⁹ is influenced by endocrine regulatory networks that during the adult stage are responsible for modulation of maternal reproductive behaviour in insects^{7,20,21}. Ovariole number is, moreover, a recognized marker of reproductive potential in the honey bee²², as well as in the well-studied solitary insect *Drosophila*^{21,23}. We found that high pollen-hoarding strain workers had more ovarioles than those from the low pollen-hoarding strain (factorial analysis of variance (ANOVA), $P < 0.005$). This difference was independent (factorial ANOVA, $P = 0.72$) of whether the workers were co-fostered (mean $\pm$ s.e.m., 5.56 ± 0.42 and 2.96 ± 0.31 ovarioles for the high and low pollen-hoarding strains, respectively; $n = 25$ per strain) or reared by their native colony (5.88 ± 0.41 and 2.88 ± 0.19 ovarioles; $n = 25$). Furthermore, bees with eight or more ovarioles were exclusively found in the high pollen-hoarding strain, where they represented 26% of the sample population (Supplementary Table S1). We also observed that this higher number of ovary filaments was associated with a swelling of the ovarioles (Supplementary Table S1), which is an established indicator of previtellogenic ovarian activation^{24,25}. These results demonstrate that a regulatory system that affects female reproductive morphology, physiology, and behaviour^{7,20,21,26} is differentially tuned during the development of honey bees characterized by different levels of pollen hoarding.

To verify that the observed variation in ovariole number translates into functional differences in adult reproductive potential, we next introduced high and low pollen-hoarding bees into host colonies with or without a queen (the presence of a queen inhibits worker oogenesis²⁷). The experimental design also controlled for rearing environment by using workers that were co-fostered and workers that were reared in their native high or low pollen-hoarding strain colony. The bees were examined after 10–21 days. In colonies with a queen ($n = 6$ colonies), we found that $29.5 \pm 3.6\%$ of the bees from the high pollen-hoarding strain ($n = 201$) had activated previtellogenic ovaries, compared with $2.6 \pm 1.8\%$ of the workers from the low pollen-hoarding strain ($n = 201$) (Supplementary Table S2). This divergence (factorial ANOVA, $P < 0.005$) was independent of whether the bees were co-fostered or reared by their native colony (factorial ANOVA, $P = 0.42$). The effect of hoarding strain on the proportion of individuals with non-activated ovaries versus previtellogenic ovaries was significant in all hives (V-square test,

¹Arizona State University, School of Life Sciences, Tempe, Arizona 85287, USA. ²University of Life Sciences, Department of Animal and Aquacultural Sciences, 1432 Aas, Norway. ³University of California, Department of Entomology, Davis, California 95616, USA.

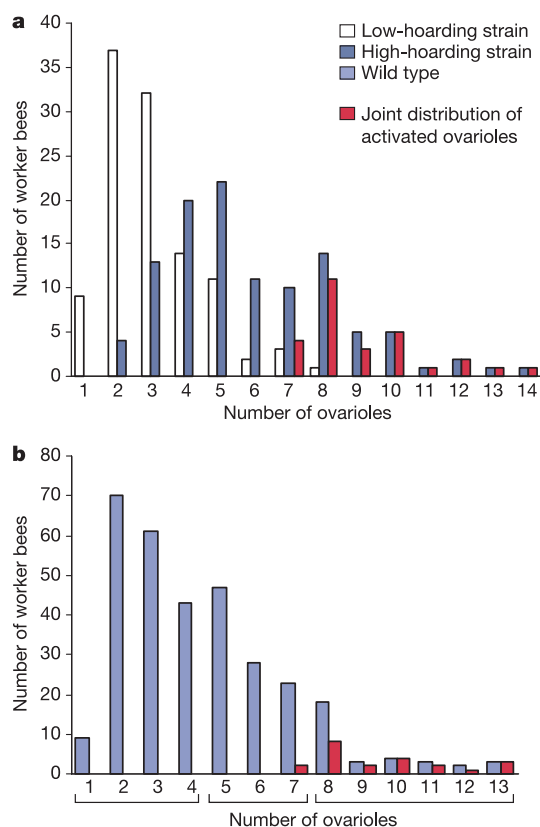


Figure 1 | Distributions of ovariole number and patterns of previtellogenic ovarian activation in worker bees. **a**, Ovariole number in mature 10- to 21-day-old bees from strains selected for high or low levels of pollen hoarding ($n = 109$ bees per strain). **b**, Samples from wild-type bees collected at presumably their first foraging flight ($n = 314$). The mean numbers of ovarioles ($\pm$ s.e.m.) for groups with 1–4, 5–7 and 8 or more ovarioles are 2.75 ± 0.06 , 5.76 ± 0.08 and 9.30 ± 0.30 , respectively. The joint distributions of ovarian activation are superimposed on the original densities and refer, therefore, to bees within the genotype-specific data sets.

$P < 0.05$). Also, previtellogenic ovarian activation was exclusively found in workers with seven or more ovarioles (Supplementary Table S2). These results from mature workers (Fig. 1a) correspond with the data from newly emerged bees (Supplementary Table S1), suggesting that a sizable proportion of worker bees selected to collect and store high amounts of pollen emerge with an active ovarian phenotype that persists for several weeks in the presence of a fully functional queen.

In colonies without a queen ($n = 6$ colonies), $75.8 \pm 0.1\%$ of the high pollen-hoarding workers ($n = 212$) had active ovaries that were previtellogenic, vitellogenic with developing oocytes, or vitellogenic with eggs (Supplementary Table S2). In comparison, $42.0 \pm 0.1\%$ of bees from the low pollen-hoarding strain had active ovaries ($n = 212$). This difference between strains (factorial ANOVA, $P < 0.05$) was independent of rearing environment (factorial ANOVA, $P = 0.94$). The effect of hoarding strain on the proportion of workers with non-activated ovaries versus previtellogenic ovaries was significant in all but one hive (V -square test, $P < 0.05$), and out of the 48 bees with eggs, 36 were from the high pollen-hoarding strain (Supplementary Table S2). Eggs were found in bees with five or more ovarioles (Supplementary Table S2). These results demonstrate that workers selected for a high level of pollen hoarding have a functional phenotype that more frequently achieves an advanced reproductive state.

Finally, we used workers from 'wild-type' colonies (not selected for pollen hoarding) to test whether the trait-associations that characterize the high and low pollen-hoarding strains are present in the general population. Wild-type bees were marked at adult eclosion and later captured at presumably their first foraging flight ($n = 551$). The nectar- and pollen-loads of the workers were quantified, and ovariole number was determined by dissection of those bees ($n = 314$) that carried measurable amounts of nectar or pollen (more than 0.0005 g).

We first investigated whether an association between ovariole number and previtellogenic ovarian activation was present. Activation occurred exclusively in bees with seven or more ovarioles (Fig. 1b), confirming our findings from the selected strains. On the basis of ovariole number, we then divided the data from the 314 workers into three groups. The first group had a mean ovariole number similar to the low pollen-hoarding strain (1–4 ovarioles, $n = 184$), the next had a mean ovariole number comparable to the high strain (5–7 ovarioles, $n = 97$), and the last group consisted of bees with eight or more ovarioles ($n = 33$) (Fig. 1b). Subsequent analysis of the data set showed that ovariole number correlated with the adult age of bees at their first foraging flight, the probability of being a pollen forager, and the nectar concentration collected by the workers (multivariate ANOVA; $P < 0.00001$). Worker bees with 5–7 and 8 or more ovarioles initiated foraging at younger ages than bees with 1–4 ovarioles (Fig. 2a). Workers with 5–7 and 8 or more ovarioles were also more likely to forage for pollen (Fig. 2b). In addition, the bees with 8 or more ovarioles collected lower nectar concentrations than workers with only 1–4 ovary filaments (Fig. 2c). Consequently, the trait-associations of wild-type bees with the greatest number of ovary filaments corresponded precisely with those shown for the strain selected to collect and store high amounts of pollen^{7,9}.

We conclude that division of foraging labour in the advanced

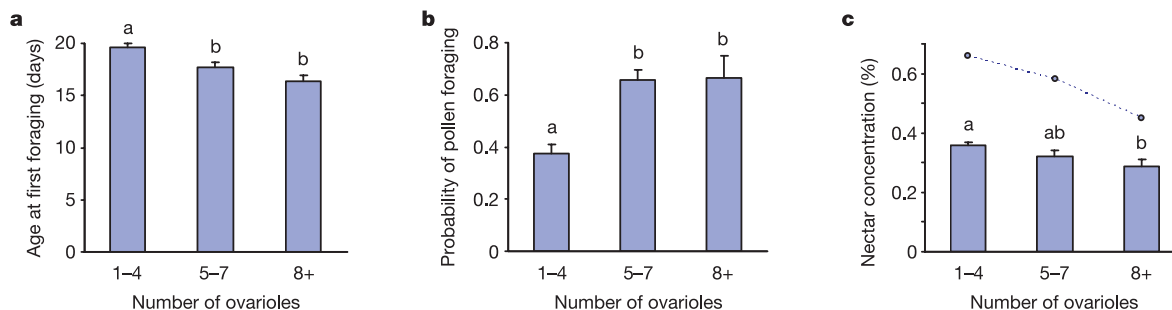


Figure 2 | Correlations between ovariole number and the social behaviour of wild-type bees. **a**, Honey bee age at presumably the first foraging flight. **b**, The probability of being a pollen forager. **c**, The sugar concentration of nectar collected by the worker bees. Data show mean $\pm$ s.e.m. Different

letters (a, b) refer to groups that were different according to a Fisher's post-hoc test ($P < 0.05$). Points connected by a dotted line in **c** denote the highest nectar concentration collected by any single bee in the respective ovariole groups.

eusocial honey bee emerges from variation in maternal care behaviour. This finding illustrates how the behavioural mechanisms of division of labour evolve from solitary ancestry, and provides an experimental demonstration of the origins of sib-care behaviour from maternal reproductive traits^{3–5,7}. The evolution of sib-care from maternal care is a critical step towards the evolution of eusociality in insects, and remains a point of substantial debate^{5,8,28,29}.

METHODS

Bees selected for high or low levels of pollen hoarding. Larvae from six high and six low pollen-hoarding strain queens were reared together in common wild-type nurse colonies. For workers reared by their native colony, frames with mature pupae were obtained from the same 12 sources. Newly emerged bees were collected for ovarian analysis or marked on the thorax with a spot of paint (Testors Enamel) for identification of strain and age. Marked workers were added to host colonies with or without a queen.

Wild-type bees. Newly emerged bees from four unrelated and unselected source/host colonies were mixed together to obtain a worker pool with high phenotypic variance. The bees were marked (see above) for identification of age, and each source/host colony received 400 workers from the mix. Starting five days later, the hive entrances were monitored between 9:00 in the morning and 14:00 in the afternoon, and marked bees that returned from flight were collected.

Foraging load measurements. Bees were treated with CO₂ until immobile to enable quantification of pollen weight, nectar weight and nectar sugar concentration, as reported previously³⁰.

Quantification of ovariole number and ovarian physiology. Bees were dissected under a stereomicroscope at ×40 magnification. Incisions were made dorsally, and the number of ovarioles in the right-side ovary²⁴ was determined at ×100 magnification. The extent of ovarian activation was determined using a relative scale as described previously²⁴: 1, non-activated ovary; 2, previtellogenic activated ovary; 3, vitellogenic ovary with developing oocytes; 4, mature ovary with at least one egg.

Data analysis. Ovariole number and ovarian activation in bees selected for high or low levels of pollen hoarding were analysed using factorial ANOVA. Analyses were combined with Fisher's post-hoc and non-parametric V-square tests to examine the effect of strain. Foraging data from wild-type bees were analysed with multivariate ANOVA and Fisher's post-hoc test. Ovariole number (coded by group: 1–4, 5–7, and 8 or more ovarioles) and host colony were the categorical factors. The effect of host colony was used to control error variance. Pollen load was coded as a binary variable. Statistica 6.0 software was used.

Received 10 September; accepted 19 October 2005.

- Robinson, G. E., Grozinger, C. M. & Whitfield, C. W. Sociogenomics: social life in molecular terms. *Nature Rev. Genet.* **6**, 257–270 (2005).
- Wilson, E. O. & Hölldobler, B. Eusociality: origin and consequences. *Proc. Natl Acad. Sci. USA* **102**, 13367–13371 (2005).
- West-Eberhard, M. J. in *Animal Societies: Theories and Facts* (eds Itô, Y., Brown, J. L. & Kikkawa, J.) 35–51 (Japan Sci. Soc. Press, Tokyo, 1987).
- West-Eberhard, M. J. in *Natural History and Evolution of Paper Wasp* (eds Turillazzi, S. & West-Eberhard, M. J.) 290–317 (Oxford Univ. Press, New York, 1996).
- Linksvayer, T. A. & Wade, M. J. The evolutionary origin and elaboration of sociality in the aculeate Hymenoptera: Maternal effects, sib-social effects, and heterochrony. *Q. Rev. Biol.* **80**, 317–336 (2005).
- West-Eberhard, M. J. *Developmental Plasticity and Evolution* (Oxford Univ. Press, New York, 2003).
- Amdam, G. V., Norberg, K., Fondrk, M. K. & Page, R. E. Reproductive ground plan may mediate colony-level selection effects on individual foraging behaviour in honey bees. *Proc. Natl Acad. Sci. USA* **101**, 11350–11355 (2004).
- Hunt, J. H. & Amdam, G. V. Bivoltinism as an antecedent to eusociality in the paper wasp genus *Polistes*. *Science* **308**, 264–267 (2005).
- Page, R. E. & Erber, J. Levels of behavioural organization and the evolution of division of labour. *Naturwissenschaften* **89**, 91–106 (2002).
- Seeley, T. D. *The Wisdom of the Hive* (Harvard Univ. Press, Cambridge, Massachusetts, 1995).
- Robinson, G. E. Regulation of division of labour in insect societies. *Annu. Rev. Entomol.* **37**, 637–665 (1992).
- Winston, M. L. *The Biology of the Honey Bee* (Harvard Univ. Press, Cambridge, Massachusetts, 1987).
- Rueppell, O., Pankiw, T. & Page, R. E. Pleiotropy, epistasis and new QTL: The genetic architecture of honey bee foraging behaviour. *J. Hered.* **95**, 481–491 (2004).
- Page, R. E. & Fondrk, M. K. The effects of colony-level selection on the social organization of honey bee (*Apis mellifera* L.) colonies: colony-level components of pollen hoarding. *Behav. Ecol. Sociobiol.* **36**, 135–144 (1995).
- Dunn, T. & Richards, M. H. When to be social: interactions among environmental constraints, incentives, guarding, and relatedness in a facultatively social carpenter bee. *Behav. Ecol.* **14**, 417–424 (2003).
- Simonet, G. et al. Neuroendocrinological and molecular aspects of insect reproduction. *J. Neuroendocrinol.* **16**, 649–659 (2004).
- Min, K. J., Taub-Montemayor, T. E., Linse, K. D., Kent, J. W. & Rankin, M. A. Relationship of adipokinetic hormone I and II to migratory propensity in the grasshopper, *Melanoplus sanguinipes*. *Arch. Insect Biochem. Physiol.* **55**, 33–42 (2004).
- Spieth, J., Nettleton, M., Zuckerapison, E., Lea, K. & Blumenthal, T. Vitellogenin motifs conserved in nematodes and vertebrates. *J. Mol. Evol.* **32**, 429–438 (1991).
- Capella, I. C. S. & Hartfelder, K. Juvenile hormone effect on DNA synthesis and apoptosis in caste-specific differentiation of the larval honey bee (*Apis mellifera* L.) ovary. *J. Insect Physiol.* **44**, 385–391 (1998).
- Tatar, M. & Yin, C. M. Slow aging during insect reproductive diapause: why butterflies, grasshoppers and flies are like worms. *Exp. Gerontol.* **36**, 723–738 (2001).
- Tu, M. P. & Tatar, M. Juvenile diet restriction and the aging and reproduction of adult *Drosophila melanogaster*. *Aging Cell* **2**, 327–333 (2003).
- Tanaka, E. D. & Hartfelder, K. The initial stages of oogenesis and their relation to differential fertility in the honey bee (*Apis mellifera*) castes. *Arthropod Struct. Dev.* **33**, 431–442 (2004).
- Hodin, J. & Riddiford, L. M. Different mechanisms underlie phenotypic plasticity and interspecific variation for a reproductive character in drosophilids (Insecta: Diptera). *Evolution* **54**, 1638–1653 (2000).
- Hartfelder, K., Bitondi, M. M. G., Santana, W. C. & Simões, Z. L. P. Ecdysteroid titer and reproduction in queens and workers of the honey bee and of a stingless bee: loss of ecdysteroid function at increasing levels of sociality? *Insect Biochem. Mol. Biol.* **32**, 211–216 (2002).
- Maurizio, A. Pollenernahrung und Lebensvorgänge bei der Honigbiene (*Apis mellifera* L.). *Landwirtsch. Jahrb. Schweiz.* **245**, 115–182 (1954).
- Hartfelder, K., Köstlin, K. & Hepperle, C. Ecdysteroid-dependent protein synthesis in caste-specific development of the larval honey bee ovary. *Roux's Arch. Dev. Biol.* **205**, 73–80 (1995).
- Butler, C. G. The control of ovary development in worker honeybees (*Apis mellifera*). *Experientia* **13**, 256–257 (1957).
- Bloch, G., Wheeler, D. & Robinson, G. E. in *Hormones, Brain and Behavior* (eds Pfaff, D., Arnold, A. P., Etgen, A. M., Fahrbach, S. E. & Rubin, R. T.) 195–235 (Academic, San Diego, 2002).
- Robinson, G. E. & Ben-Shahar, Y. Social behaviour and comparative genomics: new genes or new gene regulation? *Genes Brain Behav.* **1**, 197–203 (2002).
- Gary, N. E. & Lorenzen, K. A method for collecting the honey-sac content from honeybees. *J. Apic. Res.* **15**, 73–79 (1976).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A.L.O.T. Aase for assistance with dissections, and K. Hartfelder and P. Kukuk for comments. The project was supported by grants from the Norwegian Research Council to G.V.A., and from the National Institute on Aging and the National Research Initiative of the USDA Cooperative State Research, Education and Extension Service to R.E.P.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.V.A. (Gro.Amdam@asu.edu) or R.E.P. (Robert.Page@asu.edu).

Rapid developmental switch in the mechanisms driving early cortical columnar networks

Erwan Dupont^{1*}, Ileana L. Hanganu^{1*}, Werner Kilb¹, Silke Hirsch¹ & Heiko J. Luhmann¹

The immature cerebral cortex self-organizes into local neuronal clusters long before it is activated by patterned sensory inputs¹. In the cortical anlage of newborn mammals, neurons coassemble through electrical or chemical synapses either spontaneously^{2–4} or by activation of transmitter-gated receptors^{5,6}. The neuronal network and the cellular mechanisms underlying this cortical self-organization process during early development are not completely understood. Here we show in an intact *in vitro* preparation of the immature mouse cerebral cortex that neurons are functionally coupled in local clusters by means of propagating network oscillations in the beta frequency range. In the newborn mouse, this activity requires an intact subplate and is strongly synchronized within a cortical column by gap junctions. With the developmental disappearance of the subplate at the end of the first postnatal week⁷, activation of NMDA (*N*-methyl-D-aspartate) receptors in the immature cortical network is essential to generate this columnar activity pattern. Our findings show that during a brief developmental period the cortical network switches from a subplate-driven, gap-junction-coupled syncytium to a synaptic network acting through NMDA receptors to generate synchronized oscillatory activity, which may function as an early functional template for the development of the cortical columnar architecture.

We investigated the nature of early cortical networks in the *in vitro* intact cerebral cortex⁸ from newborn (postnatal day P0–P3) mice by recording synchronized oscillatory activity elicited by bath application of carbachol, a cholinergic agonist. Carbachol (30 μ M) elicited prominent field potential oscillations with an average maximal amplitude of 104 ± 4 μ V and a frequency of 17 ± 0.4 Hz (mean $\pm$ s.e.m., $n = 124$ recordings in 71 cortices; Fig. 1a). This oscillatory activity lasted 2–10 s, propagated at 0.15 ± 0.01 mm s^{–1} over the whole cortex, and diminished in the presence of the agonist. Similar responses could be obtained by bath application of the endogenous ligand acetylcholine (3 μ M, $n = 24$) and in seven out of nine cortices by intracortical electrical stimulation ($n = 7$; see Supplementary Fig. 1a, b).

The activity started in the caudal pole in 67 of the 90 cortices investigated and in the rostral pole in the remaining 23. Extracellular field potential recordings with a 4×4 or 1×16 electrode array revealed a tight synchronous coupling of the activity within one cortical column (coherence coefficient: intracolumnar, 0.85 ± 0.02 ; intercolumnar to neighbouring electrode, 0.73 ± 0.02 ; $P < 0.001$; both $n = 6$ cortices) and a clear distance-dependent decrease in the coherence of the activity recorded in neighbouring columns (Fig. 1b–d). These data indicate that the oscillatory activity tightly synchronizes cortical neurons within a column of 100–150 μ m in diameter.

The mechanisms underlying these synchronized oscillations were studied in a series of neuropharmacological experiments (Fig. 2a; see

also Supplementary Table 1). The network oscillations were reversibly blocked by atropine (10 μ M, $n = 13$), but not by the nicotinic cholinergic receptor antagonist mecamylamine (10 μ M, $n = 9$), indicating that muscarinic receptors are required to generate this activity. In three out of three cortices, the oscillatory activity elicited by intracortical electrical stimulation was also blocked by atropine. Neither picrotoxin (30 μ M, $n = 28$) nor SR95531 (gabazine; 100 μ M, $n = 6$) had any significant effect on the amplitude or frequency of the carbachol-induced oscillations, showing that γ -amino butyric acid A (GABA_A) receptors are not involved.

The broad spectrum glutamate receptor antagonist kynurenic acid (0.5 mM, $n = 12$) and the AMPA/kainate receptor antagonist CNQX (10 μ M, $n = 6$) did not diminish the oscillations, but blockade of NMDA receptors with CPP (10 μ M, $n = 6$) caused a significant ($P < 0.05$) decline in the amplitude by 18%. The oscillatory activity could be reversibly blocked by the connexin36-specific gap junction blocker mefloquine⁹ (25 μ M, $n = 11$) and by quinidine (100 μ M, $n = 12$), and in 6 out of 12 experiments by the broad-spectrum gap junction blocker carbenoxolone (100 μ M). In the remaining six experiments, the oscillations were reduced by 49%. In contrast to previous reports on gap-junction-dependent spontaneous neuronal domains demonstrated by Ca²⁺-imaging techniques¹⁰, the carbachol-induced oscillations were blocked by tetrodotoxin (TTX, 0.1 μ M, $n = 9$), indicating that Na⁺-dependent action potentials are required for initiating this activity.

To determine which neurons participate in these early synchronized network oscillations, we made recordings of the local field potential and whole-cell patch-clamp recordings simultaneously from different morphologically identified cell types in neocortical coronal slices of 800- μ m thickness from newborn mice. Carbachol elicited repetitive action potentials at a frequency of 3.8 ± 1.1 Hz in all cortical plate neurons ($n = 5$) and 2–3 single spikes in only two out of five layer I neurons. By contrast, all subplate neurons ($n = 12$) responded instantly to carbachol with a barrage of action potentials at a frequency of 10.8 ± 1 Hz, indicating that only subplate neurons are consistently and rapidly activated to support field potential oscillations at higher frequencies. In response to carbachol, subplate neurons showed a prominent inward current and a significant ($P < 0.001$) increase in the frequency of postsynaptic currents (PSCs) from 0.8 ± 0.2 Hz to 4 ± 0.3 Hz ($n = 44$) either before ($n = 34$) or simultaneously ($n = 13$) with the oscillatory field potential response (Fig. 3a).

Postsynaptic currents recorded during the network activity correlated with the field potential oscillation, although the coupling strength varied during the carbachol-induced response (Fig. 3b). When comparing the maximal cross-correlation coefficient between the extracellular response and the response of the cell to carbachol, layer I neurons (0.36 ± 0.01 , $n = 36$) and cortical plate neurons (0.41 ± 0.04 , $n = 6$) showed a significantly ($P < 0.05$) smaller

¹Institute of Physiology and Pathophysiology, Johannes Gutenberg University of Mainz, Duesbergweg 6, D-55128 Mainz, Germany.

*These authors contributed equally to this work.

correlation as compared with subplate neurons (0.5 ± 0.02 , $n = 36$). These data indicate that particularly subplate neurons are coupled in their activity to the local network oscillations. Further evidence for the pivotal role of subplate neurons in generating the synchronous activity came from experiments in which the subplate was removed. Whereas all cortical slabs with an intact subplate ($n = 44$) showed typical carbachol-induced oscillations (Fig. 3c), elimination of the subplate ($n = 7$) resulted in a complete loss of this activity (Fig. 3d), indicating that the subplate is necessary for the cholinergic network oscillations. Notably, in cell cultures from dissociated cortical neurons a few large GABAergic neurons, presumably subplate cells, can generate synchronous network activity¹¹.

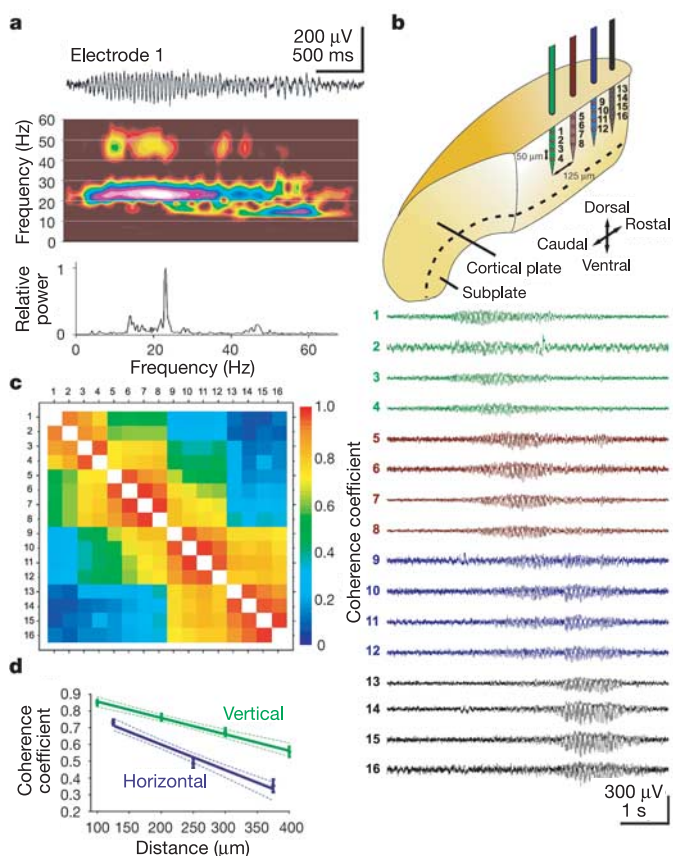


Figure 1 | Carbachol-induced network oscillations in the *in vitro* intact cerebral cortex of a P2 mouse. **a**, Top, extracellular field potential recording (electrode 1 in **b**) of the carbachol-induced activity in the somatosensory cortex recorded 50 μm below the cortical surface. Middle, colour-coded wavelet spectrum of the recording at identical timescale. Bottom, fast Fourier transformation of the recording showing maximal power at 23 Hz, as also illustrated in the wavelet spectrum. **b**, Vertical and horizontal propagation of the oscillatory activity, as recorded with a 4×4 electrode array. Top, illustration of the array. Bottom, extracellular recordings of the activity with the 16 electrodes. Responses are synchronized within a column (same colour) and progressively propagate in a horizontal direction to neighbouring recording sites. **c**, Coherence matrix of the recordings shown in **b**. Rows and columns are labelled with electrode number (1–16). Note the high coherence between recording sites located in the same column (for example, 1–4, 5–8, and so on) and the decrease in coherence between recordings in neighbouring columns (for example, 1 versus 5, 5 versus 9, and so on). **d**, Distance-dependent decrease in coherence between cholinergic oscillations recorded in the same column with a 1×16 electrode array (green, $n = 168$ coherence coefficients) and in neighbouring columns with a 4×4 electrode array (blue, $n = 126$) from P0–P3 mouse cortices (both $n = 6$). Activity recorded within the same column shows a significantly ($P < 0.001$) higher coherence than that recorded in neighbouring columns. Data are expressed as the mean $\pm$ s.e.m. and dotted lines indicate 95% confidence band of the regression lines.

Subplate neurons have several properties that put them in an ideal position to influence neuronal network activity and to assemble neighbouring neurons into a local cluster by gap junctions. First, they are well integrated in the neonatal cortical circuit and receive excitatory synaptic inputs from the thalamus, the cortical plate and other subplate neurons^{12,13}. Second, they show extensive dye-coupling after intracellular filling with biocytin (Fig. 4a and Supplementary Fig. 2), indicating that they are densely connected to other cortical neurons by gap junctions. On average, one subplate neuron is dye-coupled to 9.3 ± 1.1 other neurons in the subplate or cortical plate ($n = 42$ subplate cells, $n = 274$ coupled neurons). The average distance of the coupled neurons is $100.3 \pm 8.5 \mu\text{m}$ in the medio-lateral direction and $125.2 \pm 18.3 \mu\text{m}$ in the dorso-ventral direction. Third, they reveal a relatively high incidence of electrical coupling (15 of 89 pairs) with an average coupling conductance of $1.2 \pm 0.1 \text{ nS}$ (range 0.5–5.3 nS, $n = 15$) and a coupling ratio of $6 \pm 0.5\%$ (range 0.6–27.7%, $n = 15$; Fig. 4b–d).

Fourth, subplate neurons possess postsynaptic nicotinic¹⁴ and muscarinic receptors and receive a functional cholinergic input (Supplementary Fig. 3), which is confined to the subplate in the newborn cortex¹⁵. Current- and voltage-clamp recordings in TTX and mefloquine revealed in five out of eight subplate neurons a slow membrane depolarization of $14.8 \pm 2.5 \text{ mV}$ ($n = 5$) and an inward current of $4.5 \pm 1.5 \text{ pA}$ ($n = 3$), respectively. In 8 out of 12 subplate neurons, repetitive electrical stimulation of the cholinergic afferents elicited under blockade of glutamate and GABA_A receptors a slow and long-lasting inward current of $8.6 \pm 1.4 \text{ pA}$ in amplitude, which could be blocked by 10 μM atropine. In seven out of nine subplate cells, an atropine-sensitive slow inward current of $9.3 \pm 2.1 \text{ pA}$

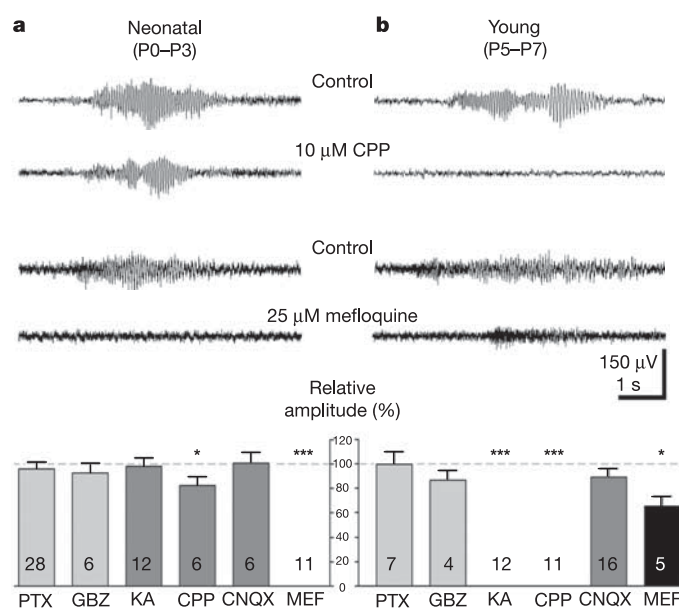


Figure 2 | Pharmacology of cholinergic oscillations in neonatal and young mouse cerebral cortex. **a**, Carbachol-induced oscillations in neonatal (P0–P3) cortex are blocked by the connexin36-specific gap junction blocker mefloquine, but not by the NMDA receptor antagonist CPP. **b**, Top, in young (P5–P7) cortex, by contrast, responses are blocked by CPP, but not by mefloquine. Bottom, effects of GABA_A, glutamate and NMDA receptor antagonists and gap junction blockers on maximal amplitude of carbachol-induced oscillations recorded with extracellular electrodes in the intact cerebral cortex of neonatal and young mice. Numbers indicate the number of recordings. Data are expressed as the mean $\pm$ s.e.m. relative to control data. Statistically significant differences are indicated at * $P < 0.05$ and *** $P < 0.001$. PTX, picrotoxin (30 μM); GBZ, gabazine (100 μM); KA, kynurenic acid (500 μM); CPP, R(–)-3-(2-carboxypiperazin-4-yl)-propyl-1-phosphonic acid (10 μM); CNQX, 6-cyano-7-nitroquinoxaline-2,3-dione (10 μM); MEF, mefloquine (25 μM).

($n = 7$) could be elicited by repetitive electrical stimulation under additional blockade of gap junctions with carbenoxolone. These data indicate that subplate neurons receive a functional cholinergic input activating postsynaptic muscarinic receptors.

Fifth, cholinergically induced PSCs recorded in subplate neurons show a pharmacological profile very similar to that of the field

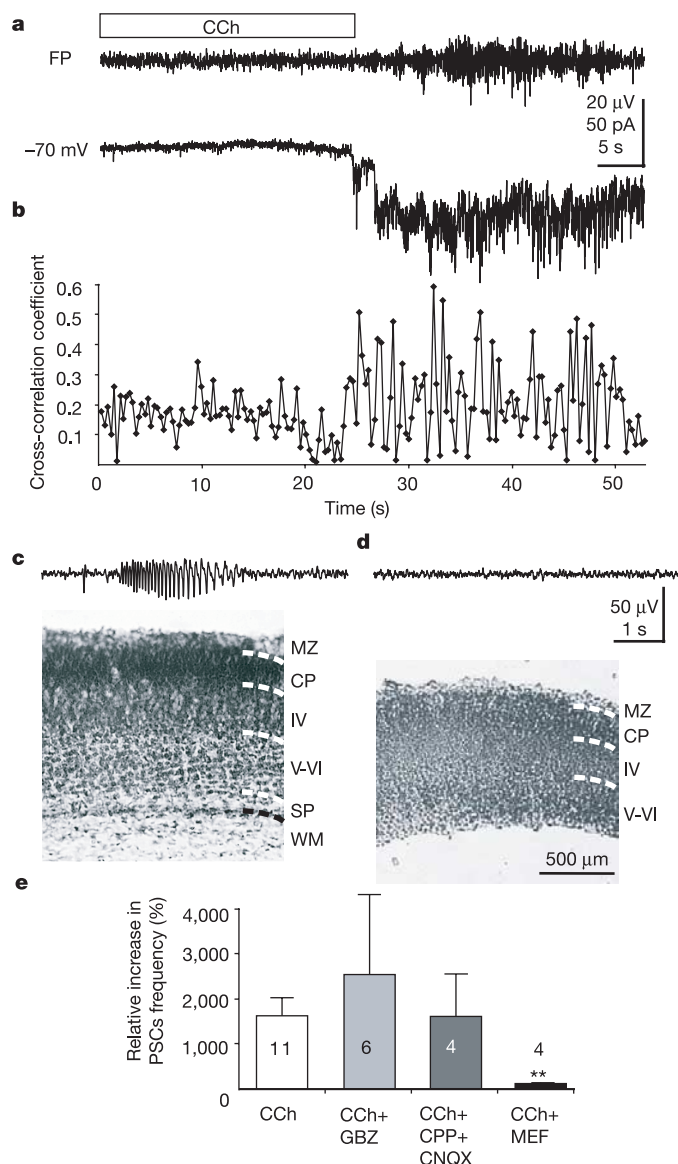


Figure 3 | Role of subplate neurons in cholinergic oscillations in the neonatal cerebral cortex. **a**, Simultaneous extracellular field potential (FP; top) and whole-cell voltage-clamp recording (bottom) from a P3 subplate neuron. Bar indicates the interval of carbachol (CCh) application. **b**, Maximal absolute cross-correlation coefficient between the extra- and intracellular response shown in **a**. Note the transient coupling and decoupling of the cell to the oscillatory field potential activity. **c**, Field potential response to carbachol application in a coronal 800- μ m thick cortical slab preparation with an intact subplate, as shown in the Nissl-stained section below. **d**, Lack of carbachol-induced activity in a cortical slab preparation in which the subplate was eliminated. Absence of the subplate was verified by Nissl staining. **e**, Relative increase in the frequency of PSCs by carbachol and effect of GABA_A, glutamate receptor or gap junction blockade on the frequency of carbachol-induced PSCs. The number of subplate neurons investigated is given for each group. Note that in the P0–P3 cortex only mefloquine (MEF) causes a significant ($P < 0.01$) decrease in the frequency of carbachol-induced PSCs, whereas blockade of GABA_A receptors with gabazine (GBZ) or glutamate receptors with CNQX plus CPP has no significant effect. Error bars are s.e.m.

potential oscillations (Fig. 3e), supporting the hypothesis that gap junctional coupling within the subplate is crucial for generating the oscillatory network activity in the neonatal cerebral cortex. To address the question of whether the subplate is also important for the spread of the synchronized oscillations, we removed different cortical layers in 800- μ m thick sagittal slices and recorded the propagation pattern (Supplementary Fig. 4). Whereas partial removal of the subplate and the lower cortical layers did not influence the spread of activity in seven out of seven slices, local isolation of the subplate by eliminating all cortical layers resulted in a propagation blockade in all seven slices investigated. These results indicate that the synchronized oscillations in newborn cortex spread in upper cortical layers.

On the basis of these data, we propose the following model for the cholinergic induction of oscillatory network activity. The cholinergic activation elicits oscillatory action potential discharges only in subplate neurons. Electrical coupling within the subplate and to the cortical plate induces, within a columnar network of 100–150 μ m in diameter, transient synchronized oscillations in the beta frequency range, as have been also demonstrated in gap-junction-coupled mature cortical networks¹⁶. Activation of muscarinic but not nicotinic receptors induces a gap-junction-mediated rise in intracellular Ca^{2+} , because only metabotropic receptors induce inositol triphosphate production. As in electrically coupled interneurons in adult cortex¹⁷, this inositol triphosphate pathway can be also stimulated by activation of metabotropic glutamate receptors, which elicits a very similar oscillatory network activity in the newborn cortex (Supplementary Fig. 1c). This activity spreads horizontally within the cortical plate through gap junctions^{18,19}. As previously reported for spontaneous Ca^{2+} waves²⁰, TTX eliminates the synchronized oscillations, indicating that action potentials augment the oscillatory activity in this network.

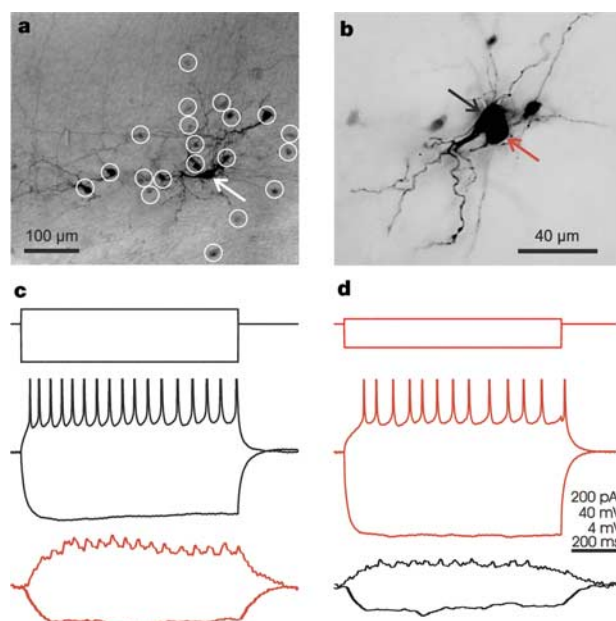


Figure 4 | Dye-coupling and electrical coupling between subplate neurons. **a**, Whole-cell recording from one P2 subplate neuron (arrow) with a biocytin-filled patch-clamp electrode results in filling of more than 15 dye-coupled neurons (circles). **b**, Photograph of two simultaneously recorded P1 subplate neurons, which were both stained with biocytin. **c**, **d**, Paired current-clamp recordings of the cells shown in **b**. Black traces were obtained from the cell marked by a black arrow in **b** and red traces from the cell marked by a red arrow. Injection of a hyperpolarizing or suprathreshold depolarizing current pulse in one cell causes voltage deflections in gap-junction-coupled neuron. Cortical surface in **a** and **b** is up; action potentials in **c** and **d** are truncated. The hyperpolarizing responses in the lower traces are the averages of four recordings.

Carbachol-induced network oscillations could be also recorded in intact cortices from older (P5–P7) mice, but these were significantly faster in frequency (22.7 ± 0.5 Hz, $n = 102$ recordings in 72 cortices, $P < 0.001$) and smaller in amplitude (92 ± 3 μ V, $P < 0.05$) as compared with those recorded in newborns. Because most subplate neurons in P5–P7 mice have been removed by programmed cell death⁷, we examined which cortical network might generate the activity in this age group (Fig. 2b and Supplementary Table 1). The oscillations in P5–P7 mice were blocked by TTX ($n = 6$) and atropine ($n = 11$), but unaffected by nicotinic ($n = 10$) and GABA_A receptor antagonists (picrotoxin, $n = 7$; gabazine, $n = 4$). In contrast to the results from P0–P3 mice and in agreement with observations in rats⁸, however, carbachol-induced oscillations in P5–P7 mice were blocked by kynurenic acid ($n = 12$) and CPP ($n = 11$), but not by CNQX ($n = 16$). These data indicate that the cholinergically induced oscillations in this age group are mediated by NMDA receptors.

Furthermore, with the developmental downregulation in gap junction coupling^{19,21}, the effects of mefloquine ($n = 5$) and quinidine ($n = 9$) on the amplitude of the oscillatory activity (reduction to 65.1% and 63.2%, respectively; both $P < 0.05$) were weaker in the P5–P7 group (Fig. 2). These data indicate that the immature cortex develops different strategies to generate synchronized network oscillations. During a brief developmental period, the network switches from a subplate-driven, gap-junction-coupled syncytium to a cortical network connected predominantly by chemical synapses containing NMDA receptors. Both networks can generate synchronized oscillatory activity in the beta frequency range and both types of activity may influence the development of cortical circuits. Ablation of subplate neurons^{22,23}, as well as early blockade of NMDA receptors^{24,25}, disturbs the development of the thalamocortical and cortical columnar organization.

So why is the immature cortex generating the same type of oscillatory activity by two distinct mechanisms and networks? The brain processes and stores information in an energy-efficient manner by binding neuronal assemblies through synchronized oscillations^{26,27}. During neonatal development—when the cortex is massively reorganized by insertion of newly generated neurons, axonal growth and synapse formation—the network relies on ontogenetically older neurons with relatively mature functional properties as the subplate cells^{12,13,28} and on local reliable interactions through electrical synapses. This network enables the immature cortex to be activated by afferent neuromodulatory inputs and to form columnar domains of gap-junction-coupled coactive neurons^{2,29}, which in the P0–P3 mouse might represent the functional correlate of cortical pre-columns before the structural formation of barrels with very similar spatial dimensions at P4 (refs 25, 30).

With the developmental disappearance of the subplate⁷ and downregulation of gap junctional coupling^{19,21}, the network uses chemical synaptic interactions through NMDA receptors to form cortical assemblies, which can be modified in an activity-dependent manner by early sensory experience. Both the gap-junction-mediated and the NMDA-receptor-dependent cortical networks rely on oscillatory interactions, because neuronal domains can be coupled very efficiently by oscillation-based synchrony even with weak electrical or chemical synaptic links²⁷, as in the immature cortex. Oscillatory activity in the beta or gamma range seems to be especially suited to create conditions in which network formation and synaptic plasticity can occur¹⁶. Our findings indicate that the subplate is pivotal in generating these intrinsic activity patterns, which may form the functional template for the experience-independent formation of early cortical networks.

METHODS

Tissue preparation. Intact cortical hemispheres and 800- μ m thick coronal slices of the primary somatosensory cortex from C57BL/6 mice aged 0–7 days were prepared as described^{8,12} in accordance with institutional guidelines. Cortical

hemispheres and slices were maintained at 32–33 °C in artificial cerebrospinal fluid containing (in mM) 124 NaCl, 26 NaHCO₃, 5 KCl, 1.6 CaCl₂, 1 MgCl₂, 1.25 NaH₂PO₄ and 10 D-glucose (pH 7.4) after equilibration with 95% O₂ and 5% CO₂.

Electrophysiology and data analysis. Extracellular field potential recordings were made in the parietal cortex with single tungsten electrodes (5 M Ω , FHC) and with a 4 $\times$ 4 or 1 $\times$ 16 channel multielectrode array (1–2 M Ω , Univ. Michigan Center for Neural Communication Technology, Ann Harbor, MI). Whole-cell recordings from single neurons or pairs of subplate cells and gramicidin-perforated patch-clamp recordings were made with fast discontinuous voltage-clamp/current-clamp amplifiers (SEC05L, npi elektronik) as described¹². In all whole-cell recordings, 0.5% biocytin was included in the patch electrode solution for morphological identification of the recorded neurons²⁸.

The extracellular carbachol-induced oscillations were analysed for their onset latency, maximal amplitude, duration and for their cross-correlation with whole-cell recordings with Spike 2 software (Cambridge Electronic Design). The spatio-temporal properties of the carbachol-induced oscillations recorded with 16 extracellular electrodes were analysed with Spid software (Vibria). We determined the coherence coefficient between all 16 recording sites and the results are illustrated in a colour-coded coherence matrix. Spontaneous PSCs were analysed with the Mini Analysis Program (Synaptosoft).

Data are presented as the mean $\pm$ s.e.m. Statistical analyses were done with Systat version 11 (SPSS Inc.) using two-tailed Student's *t*-test or Wilcoxon test. Further details are provided in the Supplementary Methods.

Received 12 July; accepted 23 September 2005.

Published online 4 December 2005.

- Katz, L. C. & Crowley, J. C. Development of cortical circuits: lessons from ocular dominance columns. *Nature Rev. Neurosci.* **3**, 34–42 (2002).
- Yuste, R., Peinado, A. & Katz, L. C. Neuronal domains in developing neocortex. *Science* **257**, 665–669 (1992).
- Garaschuk, O., Linn, J., Eilers, J. & Konnerth, A. Large-scale oscillatory calcium waves in the immature cortex. *Nature Neurosci.* **3**, 452–459 (2000).
- Khazipov, R. *et al.* Early motor activity drives spindle bursts in the developing somatosensory cortex. *Nature* **432**, 758–761 (2004).
- Peinado, A. Traveling slow waves of neural activity: a novel form of network activity in developing neocortex. *J. Neurosci.* **20**, N11–N16 (2000).
- Flint, A. C., Dammern, R. S. & Kriegstein, A. R. Endogenous activation of metabotropic glutamate receptors in neocortical development causes neuronal calcium oscillations. *Proc. Natl Acad. Sci. USA* **96**, 12144–12149 (1999).
- Price, D. J., Aslam, S., Tasker, L. & Gillies, K. Fates of the earliest generated cells in the developing murine neocortex. *J. Comp. Neurol.* **377**, 414–422 (1997).
- Kilb, W. & Luhmann, H. J. Carbachol-induced network oscillations in the intact cerebral cortex of the newborn rat. *Cereb. Cortex* **13**, 409–421 (2003).
- Cruikshank, S. J. *et al.* Potent block of Cx36 and Cx50 gap junction channels by mefloquine. *Proc. Natl Acad. Sci. USA* **101**, 12364–12369 (2004).
- Yuste, R., Nelson, D. A., Rubin, W. W. & Katz, L. C. Neuronal domains in developing neocortex: mechanisms of coactivation. *Neuron* **14**, 7–17 (1995).
- Voigt, T., Opitz, T. & De Lima, A. D. Synchronous oscillatory activity in immature cortical network is driven by GABAergic preplate neurons. *J. Neurosci.* **21**, 8895–8905 (2001).
- Hanganu, I. L., Kilb, W. & Luhmann, H. J. Functional synaptic projections onto subplate neurons in neonatal rat somatosensory cortex. *J. Neurosci.* **22**, 7165–7176 (2002).
- Friauf, E., McConnell, S. K. & Shatz, C. J. Functional synaptic circuits in the subplate during fetal and early postnatal development of cat visual cortex. *J. Neurosci.* **10**, 2601–2613 (1990).
- Hanganu, I. L. & Luhmann, H. J. Functional nicotinic acetylcholine receptors on subplate neurons in neonatal rat somatosensory cortex. *J. Neurophysiol.* **92**, 189–198 (2004).
- Mechawar, N. & Descarries, L. The cholinergic innervation develops early and rapidly in the rat cerebral cortex: A quantitative immunocytochemical study. *Neuroscience* **108**, 555–567 (2001).
- Traub, R. D., Bibbig, A., LeBeau, F. E., Buhl, E. H. & Whittington, M. A. Cellular mechanisms of neuronal population oscillations in the hippocampus *in vitro*. *Annu. Rev. Neurosci.* **27**, 247–278 (2004).
- Beierlein, M., Gibson, J. R. & Connors, B. W. A network of electrically coupled interneurons drives synchronized inhibition in neocortex. *Nature Neurosci.* **3**, 904–910 (2000).
- Kandler, K. & Katz, L. C. Coordination of neuronal activity in developing visual cortex by gap junction-mediated biochemical communication. *J. Neurosci.* **18**, 1419–1427 (1998).
- Montoro, R. J. & Yuste, R. Gap junctions in developing neocortex: a review. *Brain Res. Rev.* **47**, 216–226 (2004).
- Corlew, R., Bosma, M. M. & Moody, W. J. Spontaneous, synchronous electrical activity in neonatal mouse cortical neurons. *J. Physiol. (Lond.)* **560**, 377–390 (2004).

21. Connors, B. W., Bernardo, L. S. & Prince, D. A. Coupling between neurons of the developing rat neocortex. *J. Neurosci.* **3**, 773–782 (1983).
22. Ghosh, A. & Shatz, C. J. Involvement of subplate neurons in the formation of ocular dominance columns. *Science* **255**, 1441–1443 (1992).
23. Kanold, P. O., Kara, P., Reid, R. C. & Shatz, C. J. Role of subplate neurons in functional maturation of visual cortical columns. *Science* **301**, 521–525 (2003).
24. Fox, K., Schlaggar, B. L., Glazewski, S. & O'Leary, D. D. M. Glutamate receptor blockade at cortical synapses disrupts development of thalamocortical and columnar organization in somatosensory cortex. *Proc. Natl Acad. Sci. USA* **93**, 5584–5589 (1996).
25. Lee, L. J., Iwasato, T., Itohara, S. & Erzurumlu, R. S. Exuberant thalamocortical axon arborization in cortex-specific NMDAR1 knockout mice. *J. Comp. Neurol.* **485**, 280–292 (2005).
26. Singer, W. Development and plasticity of cortical processing architectures. *Science* **270**, 758–764 (1995).
27. Buzsáki, G. & Draguhn, A. Neuronal oscillations in cortical networks. *Science* **304**, 1926–1929 (2004).
28. Hanganu, I. L., Kilb, W. & Luhmann, H. J. Spontaneous synaptic activity of subplate neurons in neonatal rat somatosensory cortex. *Cereb. Cortex* **11**, 400–410 (2001).
29. LoTurco, J. J. & Kriegstein, A. R. Clusters of coupled neuroblasts in embryonic neocortex. *Science* **252**, 563–566 (1991).
30. Rice, F. L. & Van der Loos, H. Development of the barrels and barrel field in the somatosensory cortex of the mouse. *J. Comp. Neurol.* **171**, 545–560 (1977).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Draguhn, V. Lessmann and W. Singer for comments on the manuscript; Roche for the gift of mefloquine; and B. Krumm for technical assistance. This work was supported by grants from the Deutsche Forschungsgemeinschaft and the MAIFOR programme of the Medical Faculty at the University of Mainz.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.J.L. (luhmann@uni-mainz.de).

LETTERS

Generation of a functional mammary gland from a single stem cell

Mark Shackleton^{1,2}, François Vaillant^{1,2}, Kaylene J. Simpson^{3†}, John Stingl^{4,5}, Gordon K. Smyth¹, Marie-Liesse Asselin-Labat^{1,2}, Li Wu¹, Geoffrey J. Lindeman^{1,2} & Jane E. Visvader^{1,2}

The existence of mammary stem cells (MaSCs) has been postulated from evidence that the mammary gland can be regenerated by transplantation of epithelial fragments in mice^{1–3}. Interest in MaSCs has been further stimulated by their potential role in breast tumorigenesis⁴. However, the identity and purification of MaSCs has proved elusive owing to the lack of defined markers. We isolated discrete populations of mouse mammary cells on the basis of cell-surface markers and identified a subpopulation ($\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$) that is highly enriched for MaSCs by transplantation. Here we show that a single cell, marked with a *LacZ* transgene, can reconstitute a complete mammary gland *in vivo*. The transplanted cell contributed to both the luminal and myoepithelial lineages and generated functional lobuloalveolar units during pregnancy. The self-renewing capacity of these cells was demonstrated by serial transplantation of clonal outgrowths. In support of a potential role for MaSCs in breast cancer, the stem-cell-enriched subpopulation was expanded in premalignant mammary tissue from MMTV-*wnt-1* mice and contained a higher number of MaSCs. Our data establish that single cells within the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ population are multipotent and self-renewing, properties that define them as MaSCs.

The mammary gland is composed of a branching network of ducts and lobuloalveolar structures, the latter of which arise during pregnancy^{5,6}. Several features of the developing gland appear to require MaSCs. There are two primary epithelial cell lineages—myoepithelial and luminal (comprising ductal and alveolar subtypes)—which are presumed to arise from a common progenitor cell. The massive expansion of mammary epithelium during puberty and pregnancy, together with the remarkable regenerative capacity apparent during successive reproductive cycles, also implicate stem-like cells. Furthermore, serial transplantation of retrovirally tagged epithelial fragments has suggested that a single progenitor cell may repopulate an entire mammary gland³. Progenitor-enriched populations have been reported in the mouse mammary gland^{7,8}, and in human breast tissue, bi-potential cells capable of generating both luminal and myoepithelial cells *in vitro* have been identified^{9–11}.

Here we describe the prospective isolation of mouse MaSCs using specific cell-surface markers and performing transplantations into the mammary stroma. Because the mammary gland comprises a heterogeneous mix of cell types, we used antibodies against well-characterized endothelial (CD31) and haematopoietic (CD45 and TER119) antigens to deplete these cells using fluorescence-activated cell sorting (FACS). The substantial $\text{CD}45^+$ and $\text{CD}31^+$ populations were excluded from freshly isolated mammary cell suspensions by gating on the $\text{CD}45^- \text{CD}31^- \text{TER}119^-$ (Lin^-) population (Fig. 1a, R2). We used limiting dilution analysis¹², analogous to that employed

in the identification of haematopoietic stem cells, to determine the frequency of mammary repopulating ‘units’ (MRUs) in subpopulations of cells. Lin^- cells were isolated by FACS and transplanted in decreasing numbers into cleared mammary fat pads (MFPs) of recipient mice. The percentage of epithelial outgrowths was established for each injected cell number, and the frequency of MRUs in the Lin^- population was calculated to be 1/4,900 (Supplementary Table 1). An example of an outgrowth arising from 5,000 transplanted Lin^- cells is shown in Fig. 1b. In contrast, 22 transplants of 3,000 Lin^+ cells (Fig. 1a, R1) produced no outgrowths (Fig. 1b), indicating that MRUs are not enriched in this subset.

We defined four distinct Lin^- subpopulations based on the expression of CD29 ($\beta 1$ -integrin), a stem-cell marker in skin¹³, and CD24 (heat-stable antigen), which has been used to enrich neural stem cells¹⁴ and is expressed on human breast tumours⁴ (Fig. 1c). The frequency of MRUs in these four populations was determined following isolation by FACS and transplantation in numbers proportional to their frequency in the Lin^- population¹⁵. MRUs were enriched approximately eightfold in the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ population, whereas no significant enrichment was found in the other three subsets (Supplementary Table 2). Notably, the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ subpopulation was found to be enriched for long-term label-retaining cells, consistent with the presence of quiescent or asymmetrically dividing cells¹⁶ (Supplementary Fig. 1). In agreement with ref. 17, analysis of CD49f ($\alpha 6$ -integrin) co-expression revealed significant enrichment of $\text{CD}49^{\text{f}++}$ cells in the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ gate (data not shown). However, neither high Sca-1 (stem cell antigen) expression nor Hoechst₃₃₃₄₂ dye exclusion (which are previously reported characteristics of mammary stem/progenitor cells^{7,8}) were enriched in the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ subpopulation (Fig. 1d, e), and these observations were corroborated by transplantation of these subpopulations (Supplementary Tables 3 and 4).

We further refined the cell-purification method by double-sorting, counting and determining cell viability before transplantation. In addition, we transplanted cells from Rosa-26 mice, which carry a ubiquitously expressed *LacZ* transgene¹⁸, into wild-type recipients to allow verification of donor origin by staining for LacZ (β -galactosidase) activity in the harvested gland. Using this more quantitative method, the calculated MRU frequency in the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ population was increased to 1/64 without being significantly altered for the other populations (Supplementary Table 5). A LacZ-positive (LacZ^+) epithelial outgrowth obtained from one of these transplants is depicted in Fig. 1f. Given that cells are inevitably lost during transplantation, the actual MRU frequency is likely to be higher than 1/64.

¹The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, ²Bone Marrow Research Laboratories, Royal Melbourne Hospital, Parkville, Victoria 3050, ³Department of Biochemistry and Molecular Biology, The University of Melbourne, Victoria 3010, Australia. ⁴Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada. ⁵Stem Cell Technologies Inc., 570 West 7th Avenue, Suite 400, Vancouver, British Columbia V5Z 1B3, Canada.

[†]Present address: Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.

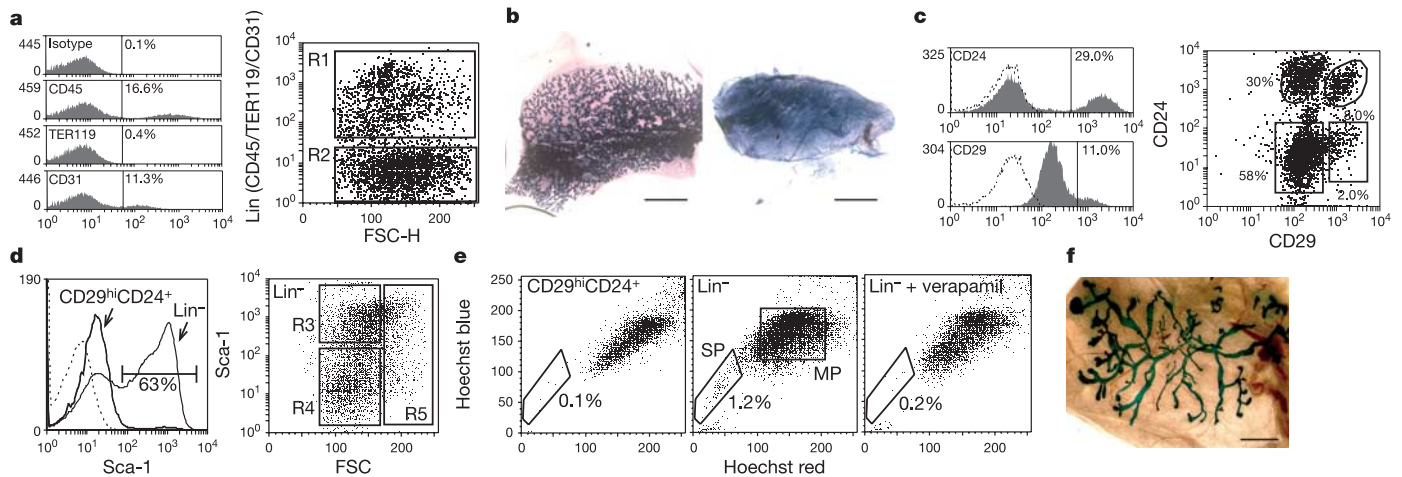


Figure 1 | Enrichment of mammary repopulating units in the $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ population. **a**, Expression of haematopoietic and endothelial cell-surface markers in mammary cell suspensions (left); gating strategy (right) used to select Lin^- (R2 gate) and Lin^+ (R1 gate) cells. **b**, Typical wholemounts of pregnant recipient MFPS transplanted with 5,000 Lin^- (left) and 3,000 Lin^+ cells (right). Scale bars in **b**, 750 μm . **c**, Expression of CD24 and CD29 in the Lin^- population (left); gating strategy used to purify cells from the four Lin^- populations (right). **d**, Expression of Sca-1 in the $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ population (left, dotted line shows isotype

labelling); gating strategy used to purify cells according to Sca-1 expression and size (right, R3–5 gates), as defined by forward scatter (FSC). Sixty-three per cent of the Lin^- population expressed high levels of Sca-1. **e**, Depletion of Hoechst side-population (SP) cells in the $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ subpopulation (left) compared to the overall Lin^- population (middle); gating strategy used to purify cells according to Hoechst staining (middle); loss of SP cells in the presence of 100 mM verapamil (right). MP, main population. **f**, A LacZ^+ outgrowth arising from transplantation of 13 visualized, double-sorted $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ cells. Scale bar in **f**, 250 μm .

To characterize epithelial cell types within the different subpopulations, flow cytometric analysis was performed using antibodies against CD24, CD29 and either the luminal marker cytokeratin 18 (K18) or the myoepithelial marker cytokeratin 14 (K14) (Fig. 2a). Whereas almost all $\text{Lin}^- \text{CD29}^{\text{lo}} \text{CD24}^+$ cells expressed K18, the majority of K14^{hi} cells resided within the $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ subpopulation and these presumably correspond to mature myoepithelial cells. Interestingly, a distinct K14^{lo} population was also revealed

in the $\text{Lin}^- \text{CD29}^{\text{lo}} \text{CD24}^+$ gate, indicating that many K18^+ cells in the mammary gland express low levels of K14.

Epithelial cell culture assays⁹ provided further evidence that the $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ population is enriched for mammary progenitor cells. Only the two $\text{Lin}^- \text{CD24}^+$ populations yielded significant colonies and the CD29^{hi} subset exhibited a two- to three-fold higher colony frequency and substantially larger colonies (Fig. 2b). To assess the differentiation capacity of the cells, we compared the growth of

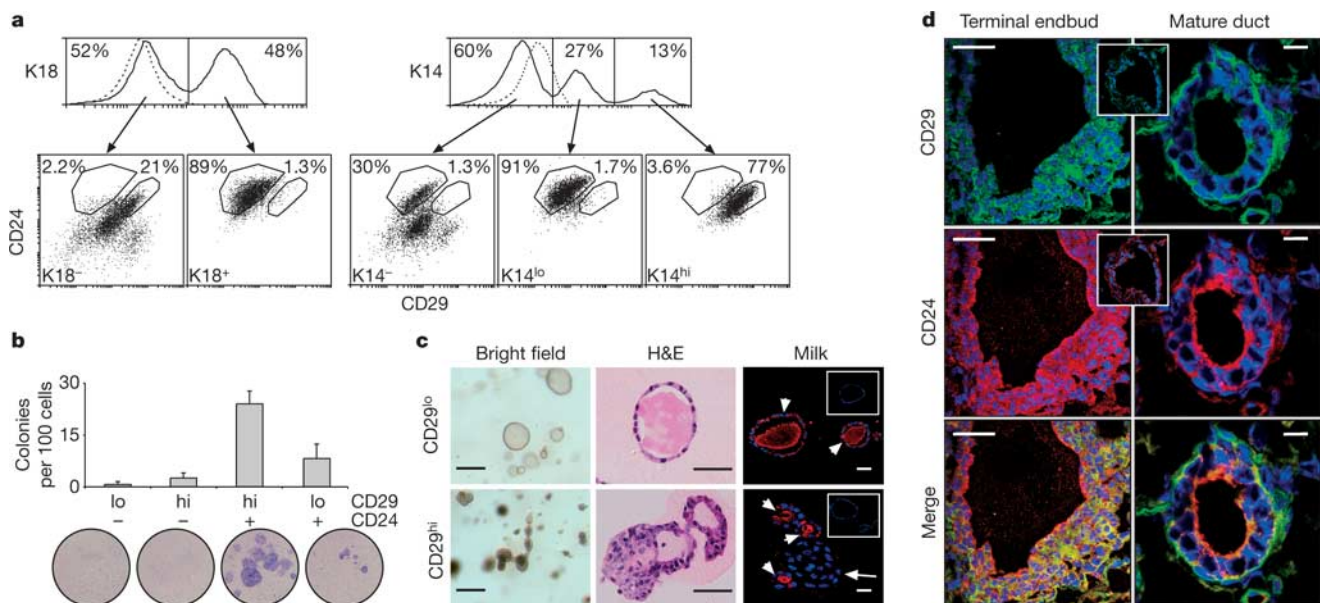


Figure 2 | Characterization of cytokeratin expression in Lin^- cells and increased progenitor capacity of $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ mammary cells *in vitro*. **a**, Expression of CD29 and CD24 in subpopulations of Lin^- cells defined by K18 and K14 expression. **b**, Colony-forming ability of the four Lin^- cell populations defined by CD29 and CD24 expression (histogram shows mean $\pm$ s.e.m., $n = 5$). **c**, Representative structures produced by Matrigel culture of $\text{Lin}^- \text{CD29}^{\text{lo}} \text{CD24}^+$ (upper row) and $\text{Lin}^- \text{CD29}^{\text{hi}} \text{CD24}^+$ cells (lower row); bright-field views of gels (left; scale bars, 100 μm), H&E-stained sections (middle; scale bars, 10 μm), and labelling with

anti-milk antibody are shown (right; arrowheads indicate milk-producing structures, arrow indicates a non-milk-producing structure). Insets in right panels show isotype-labelled control sections: red, milk; blue, DAPI; scale bars, top 40 μm , bottom 20 μm . **d**, Expression of CD29 and CD24 in a terminal endbud (left column; scale bar, 40 μm) and a mature duct (right column; scale bar, 16 μm). The merged confocal images are shown below. Insets show isotype-labelled control sections: green, CD29; red, CD24; blue, DAPI.

Lin⁻CD29^{hi}CD24⁺ and Lin⁻CD29^{lo}CD24⁺ cells in Matrigel under lactogenic conditions. Cells from the Lin⁻CD29^{lo}CD24⁺ population predominantly (85%) formed single-cell layered, alveolar-like structures that produced milk protein (Fig. 2c and Supplementary Fig. 2a). Thus, this population appears to be enriched for progenitors that have a luminal cell fate under lactogenic conditions. In contrast, Lin⁻CD29^{hi}CD24⁺ cells formed a heterogeneous mix of morphologically distinguishable structures, including ductal forms and multicellular spheroid bodies, as well as occasional (3%) alveolar-like structures akin to those from the Lin⁻CD29^{lo}CD24⁺ population (Fig. 2c and Supplementary Fig. 2a). Co-staining of the Matrigel-derived structures with antibodies to K14 and K18 revealed marked differences between Lin⁻CD29^{lo}CD24⁺ and Lin⁻CD29^{hi}CD24⁺ cell-derived structures with respect to their pattern of cytokeratin expression (Supplementary Fig. 2b). The expanded differentiative repertoire and enhanced colony-forming ability of Lin⁻CD29^{hi}CD24⁺ cells suggest that this population is enriched for mammary progenitors. Compatible with these findings, high levels of CD29 expression were apparent in the cap cell region of terminal end buds, presumed to be rich in stem cells^{19,20}, relative to mature ducts in which high expression was predominantly basolateral, as previously shown²¹ (Fig. 2d).

To test the 'common-progenitor model' of lineage development in the mammary gland, we determined whether the Lin⁻CD29^{hi}CD24⁺ MRU constituted a single cell. Double-sorted Lin⁻CD29^{hi}CD24⁺ cells from Rosa-26 mice were resuspended at a concentration of one cell per injection volume, with or without supporting cells (5,000 cells from a wild-type population depleted of Lin⁻CD29^{hi}CD24⁺ cells). Eight LacZ⁺ epithelial outgrowths were produced from 68 injections (Supplementary Table 6). Notably, supporting cells did not affect the likelihood of an outgrowth or its size. Although the eight outgrowths could have resulted from more than one distinct progenitor, calculations showed this to be extremely unlikely (see Supplementary Methods). To prove definitively that a single cell could repopulate a cleared MFP, we transplanted individual, double-sorted Lin⁻CD29^{hi}CD24⁺ Rosa-26 cells that had been viewed microscopically in 10- μ l Terasaki wells. Six LacZ⁺ outgrowths were produced from 102 transplants in three independent experiments (Supplementary Table 6 and Fig. 3a) and, as previously observed, supporting cells had no effect. Substantial engraftment of the fat pad was evident and histological sectioning of the outgrowths revealed normal ductal structures composed of both luminal and myoepithelial cells (Fig. 3b). Furthermore, repopulated glands exhibited complete functional differentiation at parturition (Fig. 3a) and sections derived from pregnant recipients revealed lipid droplets and abundant milk protein within alveoli and ductal lumens, respectively (Fig. 3b and c). Thus, a single Lin⁻CD29^{hi}CD24⁺ cell can reconstitute a completely functional mammary gland, demonstrating its high proliferative and multi-lineage differentiation capacity.

Further evidence for the clonality of outgrowths arising from limiting numbers of Lin⁻CD29^{hi}CD24⁺ cells was derived from 'mixing' experiments in which cells from wild-type and Rosa-26 mice were co-injected. Seventy cells from each donor were mixed before transplantation and outgrowths analysed in virgin recipients. Pure wild-type or LacZ⁺ outgrowths were observed in 95/97 cases (Supplementary Fig. 3), indicating that combinatorial activity between independent MaSCs is not essential to generate a mammary outgrowth.

To evaluate whether the Lin⁻CD29^{hi}CD24⁺ repopulating cell can self-renew, epithelial outgrowths derived from primary transplants of Lin⁻CD29^{hi}CD24⁺ cells were analysed by flow cytometry and serial transplantation. The primary outgrowths comprised the same CD29 and CD24 profiles as donor mice (Fig. 3d, left panel), whereas cell suspensions from cleared, untransplanted mammary fat pads were CD24⁻ (Fig. 3d, right panel), demonstrating that the CD24⁺ cells were donor-derived. For secondary transplantation, we used primary outgrowths that developed from up to 25

Lin⁻CD29^{hi}CD24⁺ Rosa-26 cells and which were therefore likely to be clonal (see Supplementary Methods). Cells from each primary outgrowth generated LacZ⁺ outgrowths in multiple secondary recipients (Supplementary Table 7 and Supplementary Methods), indicating that self-renewal had occurred in the primary outgrowths. Tertiary transplantations were also performed and full developmental

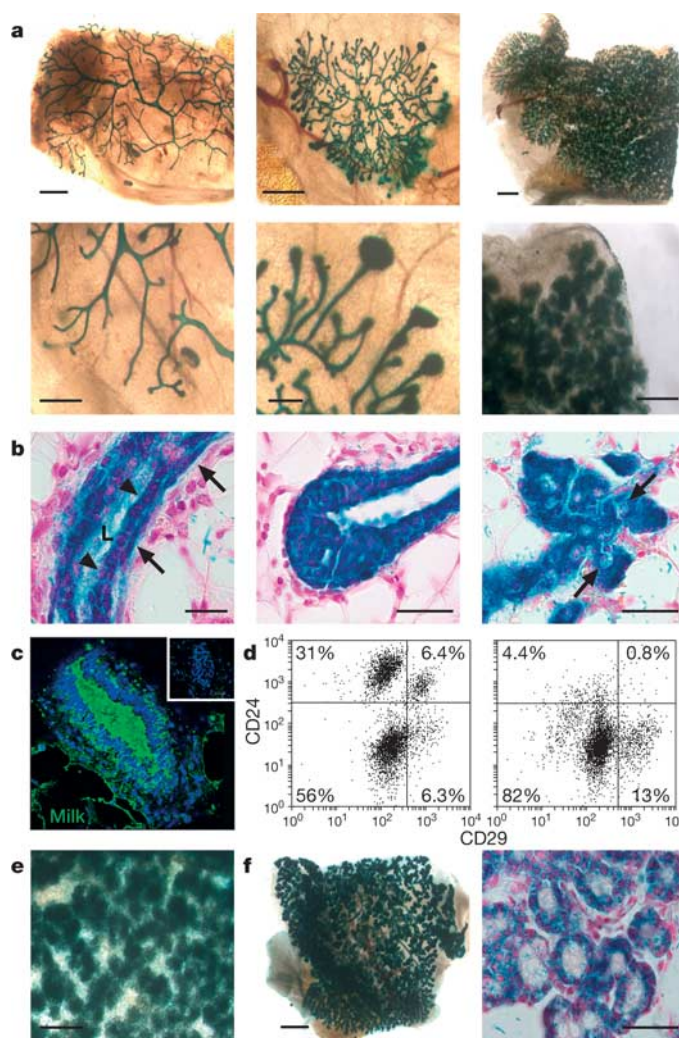


Figure 3 | A single, self-renewing Lin⁻CD29^{hi}CD24⁺ cell can repopulate a MFP. **a**, Wholemount analysis of epithelial outgrowths arising from transplantation of a single LacZ⁺ Lin⁻CD29^{hi}CD24⁺ cell; virgin recipients harvested 10 and 8.5 weeks after transplantation (top left and middle; scale bars, 250 μ m), and a day-19 pregnant recipient (top right; scale bar, 250 μ m). High-magnification images of ductal structures (bottom left; scale bar, 100 μ m), terminal endbuds (bottom middle; scale bar, 50 μ m), and lobulo-alveolar structures in a full-term pregnant recipient (bottom right; scale bar, 100 μ m). **b**, Sections of single-cell origin, LacZ⁺ outgrowths stained with nuclear fast red show ductal (arrowheads) and myoepithelial (arrows) cell lineages (left, scale bar, 5 μ m) and a terminal endbud (middle; scale bar, 10 μ m) in a virgin recipient, and lobulo-alveolar epithelium in a mid-term pregnant recipient (right, arrows indicate lipid droplets; scale bar, 10 μ m). L, lumen. **c**, Immunofluorescence staining with anti-milk antibody of a duct arising from a single LacZ⁺ Lin⁻CD29^{hi}CD24⁺ cell in a pregnant recipient; inset, isotype-labelled control section; green, milk; blue, DAPI. **d**, Flow cytometric analysis of cell suspensions prepared from MFPs transplanted with Lin⁻CD29^{hi}CD24⁺ cells (left) and untransplanted cleared MFPs (right). **e**, A LacZ⁺ outgrowth at parturition that arose from transplantation of cells derived from an outgrowth of single Lin⁻CD29^{hi}CD24⁺ cell origin; scale bar, 100 μ m. **f**, A tertiary LacZ⁺ outgrowth harvested at parturition (left; scale bar, 250 μ m). Section of this outgrowth stained with nuclear fast red (right; scale bar, 10 μ m).

capacity was retained in these outgrowths (Fig. 3f). To definitively confirm the self-renewing capability of the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ repopulating cell, we transplanted a primary outgrowth of single-cell origin and produced 15 secondary outgrowths (Supplementary Table 7) that exhibited complete differentiation at parturition (Fig. 3e). Thus, the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ mammary repopulating cell is capable of self-renewal, a defining feature of stem cells²².

Recent evidence suggests the existence of a tumour stem cell for breast cancer⁴. We therefore examined expression of CD29 and CD24 in hyperplastic but premalignant mammary tissue from two strains of mice prone to develop tumours. Despite an age-related expansion of the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ subpopulation in wild-type glands, significantly increased epithelial cellularity and percentage of $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ cells were evident in MMTV-*wnt-1* mice (Fig. 4a and b). In addition, transplantation studies showed the relative MaSC frequencies in $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ subpopulations from MMTV-*wnt-1* and control mice to be 1/57 (1/37 to 1/86) and 1/86 (1/51 to 1/147), respectively. Collectively, these data indicated a 6.4-fold (± 1.2 s.e.m.) increase in the absolute number of MaSCs in premalignant MMTV-*wnt-1* transgenic glands, implying a role for Wnt signalling in the self-renewal of MaSCs analogous to its role in haematopoietic stem cells²³. In addition, the epithelial outgrowths arising from transplantation of $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ MMTV-*wnt-1* mammary cells were profoundly hyperplastic at 5 weeks post-transplantation (Fig. 4c). Our findings are compatible with the proposal that the *wnt-1* oncogene gives rise to heterogeneous

tumours because it perturbs the MaSC pool^{24,25}. Interestingly, pre-neoplastic mammary tissue from MMTV-*neu* mice²⁶, which develop luminal epithelial tumours, showed no expansion of the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ population (Fig. 4a and b). Taken together, our data suggest that different epithelial cell-types are the targets of transformation in the *wnt-1* and *neu* mammary tumorigenesis models.

This study provides the first description, to our knowledge, of the reconstitution of an entire organ from a single epithelial stem cell and has implications for the isolation of stem cells from other epithelial tissues. It is not known whether there is a hierarchy of mammary stem and progenitor cells, analogous to the haematopoietic system^{22,27}. However, there is evidence suggesting that distinct cellular progenitors for ductal and lobular structures exist in the mammary gland^{3,28}. We have established that a single stem cell is capable of reconstituting the entire mammary tree. Presumably this cell undergoes division in its stromal milieu to yield progenitors as well as daughter stem cells that cooperate to direct coordinated expansion of mammary epithelium into a branching ductal network. It seems likely that $\beta 1$ -integrin and $\alpha 6$ -integrin (ref. 17) participate in mediating interactions between MaSCs and the mammary stroma. Delineation of genes that are expressed in stem and progenitor cells should allow the identification of further markers of the MaSC and the putative breast cancer stem cell.

METHODS

Mice. FVB/NJ, C57BL/6, BALB/c, Rosa-26 (ref. 18) (C57BL/6), MMTV-*wnt-1* (BALB/c), and MMTV-*neu* (FVB/NJ) mice were bred and maintained in our animal facility according to institutional guidelines. All experiments were approved by the Animal Research Ethics Committee of the Melbourne Health Research Directorate.

Mammary cell preparation. Mammary glands were dissected from 8-week-old female mice. After mechanical dissociation with a McIlwain tissue chopper (Mickle Laboratory Engineering), the tissue was placed in culture medium (DME HAM with 1 mM glutamine, 5 $\mu\text{g ml}^{-1}$ insulin, 500 ng ml^{-1} hydrocortisone, 10 ng ml^{-1} epidermal growth factor and 20 ng ml^{-1} cholera toxin) supplemented with 5% bovine calf serum and containing 300 U ml^{-1} collagenase (Sigma) and 100 U ml^{-1} hyaluronidase (Sigma), and digested for 1 h at 37 °C. The resultant organoid suspension was sequentially resuspended in 0.25% trypsin-EGTA for 1–2 min, 5 mg ml^{-1} dispase (Roche Diagnostics) and 0.1 mg ml^{-1} DNase (Worthington) for 5 min, and 0.64% NH_4Cl for 3 min before filtration through a 40- μm mesh and labelling.

Antibodies. Antibodies against mouse antigens were purchased from BD Pharmingen unless otherwise specified, and included CD24-PE, biotinylated and APC-conjugated CD31, biotinylated and APC-conjugated CD45, biotinylated TER119, Sca-1-FITC and -PE, biotinylated CD49f, CD29-FITC (Chemicon), anti-milk (Nordic Immunological Laboratories), anti-cytokeratin 14 (Covance), and anti-cytokeratin 18 (Progen Biotechnik). Streptavidin-APC was purchased from BD Pharmingen. Fluorochrome-conjugated secondary antibodies included anti-rabbit and anti-mouse Ig-Alexa₅₉₄ and anti-rabbit Ig-Alexa₄₈₈ (Molecular Probes).

Cell labelling, flow cytometry and sorting. Hoechst staining was performed for 1 h at 37 °C with 6 $\mu\text{g ml}^{-1}$ Hoechst₃₃₃₄₂ (Sigma). For labelling of intracellular epitopes, cells were fixed in chilled acetone for 1 min and permeabilized in 0.1% Tween/PBS on ice for 5 min before blocking. Blocking was performed in rat γ -globulin (Jackson Laboratories) and anti-CD16/CD32 Fc γ III/II receptor antibody (BD Pharmingen) for 10 min. Antibody incubations were performed at 4 °C for 25 min (45 min for intracellular epitope labelling). Cells were resuspended in 0.5 $\mu\text{g ml}^{-1}$ propidium iodide (Sigma) before analysis. Data analysis was performed on the single, live cell gate using WEASEL software (<http://www.wehi.edu.au/cytometry/WEASEL2.html>). Cell sorting was carried out on a FACSDiVa, FACStar or FACS Vantage cell sorter (Becton Dickinson). The purity of sorted populations was routinely more than 95%.

Mammary fat pad transplantation and analysis. Sorted cells were resuspended in PBS with 0.04% trypan blue (Sigma) and 50% fetal calf serum (FCS), and injected in 10 μl volumes into the inguinal glands of 3-week-old female mice that had been cleared of endogenous epithelium¹. Visualization of cells before transplantation was performed in 10- μl Terasaki wells. Recipient glands were removed for evaluation after 5–10 weeks unless otherwise stated. Wholemouts of wild-type mammary outgrowths were stained with haematoxylin. LacZ⁺ outgrowths were detected by X-Gal staining²⁹, and/or polymerase chain reaction analysis for the *LacZ* gene¹⁸. An outgrowth was defined as an epithelial structure

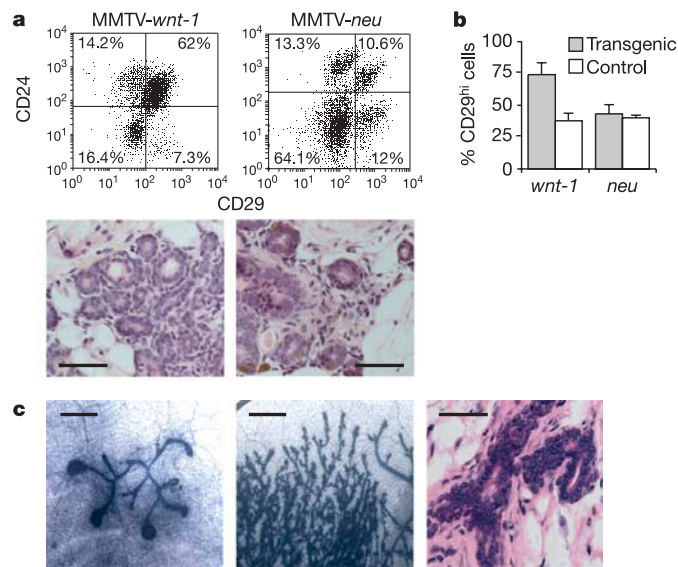


Figure 4 | The $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ population is expanded in MMTV-*wnt-1* transgenic mice. **a**, Flow cytometric analyses of CD24 and CD29 expression in cell suspensions from preneoplastic MMTV-*wnt-1* and MMTV-*neu* transgenic mammary glands. Tissue was taken from parous MMTV-*wnt-1* mice ($n = 3$) at 4 months of age and virgin MMTV-*neu* mice ($n = 3$) at 6 months of age. Also shown are representative H&E-stained sections from the same premalignant glands (scale bars, 40 μm). **b**, Histogram depicting the percentages (mean $\pm$ s.e.m.) of CD29^{hi} cells in the $\text{Lin}^- \text{CD}24^+$ populations of MMTV-*wnt-1* ($n = 3$; 74%) and MMTV-*neu* ($n = 3$; 43%) mammary glands compared with age- and parity-matched controls ($n = 2$; 38% and 40%, respectively). A similar fold expansion of the $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ subpopulation was observed in nulliparous and parous MMTV-*wnt-1* mice. However, this population was found to increase with age in wild-type mice (1.5- to twofold larger in 4–6-month-old mice versus 8-week-old mice). **c**, Wholemounts of outgrowths arising in virgin recipient mice 5 weeks post-transplantation of 25 $\text{Lin}^- \text{CD}29^{\text{hi}} \text{CD}24^+$ cells derived from control (left) or MMTV-*wnt-1* (central) donor mice; scale bars, 250 μm . Right panel shows an H&E-stained section prepared from the structure depicted in the middle panel. Scale bars, 20 μm .

comprising ducts arising from a central point, with lobules and/or terminal end buds.

In vitro assays. For colony assays, cells were sorted directly into the wells of 24-well plates containing culture medium with 1% FCS in the presence of $14,000\text{ cm}^{-2}$ irradiated NIH-3T3 cells. After 24 h the media was replaced with serum-free culture medium containing 0.1% bovine serum albumin, and 5 days later the colonies were fixed with acetone:methanol (1:1), stained with Giemsa, and counted. For three-dimensional assays, cells were resuspended in chilled 100% Matrigel and the gels allowed to set before covering with culture medium supplemented with 1% FCS. To induce differentiation, the medium was changed to DME-HAM containing 1 mM glutamine, $5\text{ }\mu\text{g ml}^{-1}$ insulin, 500 ng ml^{-1} hydrocortisone, $5\text{ }\mu\text{g ml}^{-1}$ prolactin and 1% FCS after 1 week. The cells were cultured for a total of 2 weeks before fixation in 4% paraformaldehyde and embedding in paraffin for sectioning and staining with haematoxylin and eosin (H&E), or fixation in chilled acetone:methanol (1:1) for immunostaining.

Immunostaining. Frozen sections were prepared from tissues embedded in OCT compound. After fixation in 100% acetone, sections were rehydrated and blocked with 5% bovine calf serum in PBS. Paraffin-embedded sections were dewaxed, washed in PBS, and subjected to antigen retrieval by boiling in 10 mM citrate buffer for 20 min and treatment with 150 mM glycine for 15 min, before blocking as above. Primary antibody staining was performed overnight at 4°C , while secondary antibody staining was performed for 30 min and DAPI staining for 5 min at room temperature. Matrigel cultures were prepared and immunostained as described³⁰. Sections were imaged on a Leica TCS4 SP2 spectral confocal scanner linked to a Leica DMIRE2 inverted microscope.

Received 1 August; accepted 13 October 2005.

- DeOme, K. B., Faulkin, L. J. Jr, Bern, H. A. & Blair, P. B. Development of mammary tumors from hyperplastic alveolar nodules transplanted into gland-free mammary fat pads of female C3H mice. *Cancer Res.* **19**, 515–520 (1959).
- Daniel, C. W., DeOme, K. B., Young, J. T., Blair, P. B. & Faulkin, L. J. Jr. The *in vivo* life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *Proc. Natl Acad. Sci. USA* **61**, 53–60 (1968).
- Kordon, E. C. & Smith, G. H. An entire functional mammary gland may comprise the progeny from a single cell. *Development* **125**, 1921–1930 (1998).
- Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
- Hennighausen, L. & Robinson, G. W. Think globally, act locally: the making of a mouse mammary gland. *Genes Dev.* **12**, 449–455 (1998).
- Daniel, C. W. & Smith, G. H. The mammary gland: a model for development. *J. Mammary Gland Biol. Neoplasia* **4**, 3–8 (1999).
- Welm, B. E. *et al.* Sca-1^{pos} cells in the mouse mammary gland represent an enriched progenitor cell population. *Dev. Biol.* **245**, 42–56 (2002).
- Alvi, A. J. *et al.* Functional and molecular characterisation of mammary side population cells. *Breast Cancer Res.* **5**, R1–R8 (2003).
- Stingl, J., Eaves, C. J., Zandieh, I. & Emsman, J. T. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res. Treat.* **67**, 93–109 (2001).
- Gudjonsson, T. *et al.* Isolation, immortalization, and characterization of a human breast epithelial cell line with stem cell properties. *Genes Dev.* **16**, 693–706 (2002).
- Dontu, G. *et al.* *In vitro* propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes Dev.* **17**, 1253–1270 (2003).
- Bonnefoix, T., Bonnefoix, P., Verdiel, P. & Sotto, J. J. Fitting limiting dilution experiments with generalized linear models results in a test of the single-hit Poisson assumption. *J. Immunol. Methods* **194**, 113–119 (1996).
- Jones, P. H., Harper, S. & Watt, F. M. Stem cell patterning and fate in human epidermis. *Cell* **80**, 83–93 (1995).
- Rietze, R. L. *et al.* Purification of a pluripotent neural stem cell from the adult mouse brain. *Nature* **412**, 736–739 (2001).
- Uchida, N. & Weissman, I. L. Searching for hematopoietic stem cells: evidence that Thy-1^{lo} Lin[−] Sca-1⁺ cells are the only stem cells in C57BL/Ka-Thy-1.1 bone marrow. *J. Exp. Med.* **175**, 175–184 (1992).
- Smith, G. H. Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* **132**, 681–687 (2005).
- Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* advance online publication, 4 January 2006 (doi:10.1038/nature04496).
- Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5**, 1513–1523 (1991).
- Kenney, N. J., Smith, G. H., Lawrence, E., Barrett, J. C. & Salomon, D. S. Identification of stem cell units in the terminal end bud and duct of the mouse mammary gland. *J. Biomed. Biotechnol.* **1**, 133–143 (2001).
- Williams, J. M. & Daniel, C. W. Mammary ductal elongation: differentiation of myoepithelium and basal lamina during branching morphogenesis. *Dev. Biol.* **97**, 274–290 (1983).
- White, D. E. *et al.* Targeted disruption of $\beta 1$ -integrin in a transgenic mouse model of human breast cancer reveals an essential role in mammary tumour induction. *Cancer Cell* **6**, 159–170 (2004).
- Weissman, I. L. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* **100**, 157–168 (2000).
- Reya, T. *et al.* A role for Wnt signalling in self-renewal of haematopoietic stem cells. *Nature* **423**, 409–414 (2003).
- Li, Y. *et al.* Evidence that transgenes encoding components of the Wnt signalling pathway preferentially induce mammary cancers from progenitor cells. *Proc. Natl Acad. Sci. USA* **100**, 15853–15858 (2003).
- Liu, B. Y., McDermott, S. P., Khwaja, S. S. & Alexander, C. M. The transforming activity of Wnt effectors correlates with their ability to induce the accumulation of mammary progenitor cells. *Proc. Natl Acad. Sci. USA* **101**, 4158–4163 (2004).
- Guy, C. T. *et al.* Expression of the *neu* protooncogene in the mammary epithelium of transgenic mice induces metastatic disease. *Proc. Natl Acad. Sci. USA* **89**, 10578–10582 (1992).
- Dick, J. E. Stem cells: Self-renewal writ in blood. *Nature* **423**, 231–233 (2003).
- Smith, G. H. Experimental mammary epithelial morphogenesis in an *in vivo* model: evidence for distinct cellular progenitors of the ductal and lobular phenotype. *Breast Cancer Res. Treat.* **39**, 21–31 (1996).
- Sum, E. Y. M., O'Reilly, L. A., Jonas, N., Lindeman, G. J. & Visvader, J. E. The LIM domain protein Lmo4 is highly expressed in proliferating mouse epithelial tissues. *J. Histochem. Cytochem.* **53**, 475–486 (2005).
- Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**, 256–268 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to N. Forrest for expert assistance, J. Adams and A. Harris for critical review of the manuscript, S. Mihajlovic for histology, and F. Battye and A. Holloway for FACS support. This work was supported by the Victorian Breast Cancer Research Consortium and the National Health and Medical Research Council (Australia). M.-L.A.-L. is supported by a Fondation pour la Recherche Médicale Fellowship, K.J.S. by a Peter Doherty Fellowship and J.S. by Fellowships from the Canadian Breast Cancer Foundation and the Natural Sciences and Engineering Research Council of Canada.

Author Contributions M.S. and F.V. contributed equally to this work. G.J.L. and J.E.V. contributed equally to this work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.E.V. (visvader@wehi.edu.au) or G.J.L. (lindeman@wehi.edu.au).

An siRNA-based microbicide protects mice from lethal herpes simplex virus 2 infection

Deborah Palliser^{1,2}, Dipanjan Chowdhury^{1,2}, Qing-Yin Wang³, Sandra J. Lee⁴, Roderick T. Bronson⁵, David M. Knipe³ & Judy Lieberman^{1,2}

Herpes simplex virus 2 (HSV-2) infection causes significant morbidity¹ and is an important cofactor for the transmission of HIV infection². A microbicide to prevent sexual transmission of HSV-2 would contribute substantially to controlling the spread of HIV and other infections^{3,4}. Because RNA interference (RNAi) provides effective antiviral defence in plants and other organisms, several studies have focused on harnessing RNAi to inhibit viral infection⁵. Here we show that vaginal instillation of small interfering RNAs (siRNAs) targeting HSV-2 protects mice from lethal infection. siRNAs mixed with lipid are efficiently taken up by epithelial and lamina propria cells and silence gene expression in the mouse vagina and ectocervix for at least nine days. Intravaginal application of siRNAs targeting the HSV-2 *UL27* and *UL29* genes (which encode an envelope glycoprotein and a DNA binding protein⁶, respectively) was well tolerated, did not induce interferon-responsive genes or cause inflammation, and protected mice when administered before and/or after lethal HSV-2 challenge. These results suggest that siRNAs are attractive candidates for the active component of a microbicide designed to prevent viral infection or transmission.

Most mammalian cells do not take up siRNAs without a transfection reagent. We instilled fluorescein isothiocyanate (FITC)-labelled siRNAs complexed with a transfection lipid into the mouse vagina. The vaginal and ectocervical epithelium, underlying lamina propria and stroma efficiently took up the fluorescent siRNAs (Fig. 1a). When siRNAs targeting enhanced green fluorescent protein (*EGFP*) were administered intravaginally with lipid to transgenic GFP mice that express *EGFP* in every cell from the β -actin promoter⁷, GFP expression three days later was down-modulated throughout the vagina and cervix of GFP siRNA-treated mice, but not in control mice (Fig. 1b). Intravaginal siRNAs did not cause systemic silencing in distant organs such as the liver. Silencing persisted without diminution for at least nine days (the total length of the experiments) under conditions in which epithelial turnover was reduced by treatment with medroxyprogesterone acetate (Fig. 1c). Further studies are required to determine how long silencing persists and to assess the effect of menstrual variation on durability. Nonetheless, the extent and persistence of silencing suggests that siRNAs are attractive candidates for the active component of a microbicide. Moreover, their durability suggests that an RNAi-based microbicide might not need to be administered just before sexual intercourse, mitigating one of the main problems with microbicides: compliance.

To determine whether topical siRNA application could protect against sexually transmitted infection, seven siRNAs targeting three essential HSV-2 genes—*UL5* (a component of the helicase–primase complex), *UL27* (envelope glycoprotein B) and *UL29* (a DNA-binding protein)⁶—were designed using the Dharmacon design program⁸. After overnight incubation, siRNA-treated NIH3T3

(Fig. 2a) and Vero (Fig. 2b) cells were infected with HSV-2 strain 186 at a multiplicity of infection (MOI) of 1, and viral replication was assessed by plaque assay 24 h later. *UL5.2*, *UL27.2* and *UL29.2* siRNAs significantly reduced viral titre, but GFP siRNA and inverted *UL29.2* siRNA did not (Fig. 2b, c).

UL29.2 was the most effective siRNA, suppressing viral replication by 62-fold in NIH3T3 cells and 25-fold in Vero cells. Viral replication by *UL29.2* was inhibited at siRNA concentrations of 25 nM, and reached a plateau at 100 nM siRNA (Fig. 2c and data not shown). Gene silencing was specific for the targeted gene. When *UL27* and *UL29* messenger RNAs were quantified by real-time polymerase chain reaction with reverse transcription (RT-PCR) in Vero cells transfected one day earlier with *UL27.2*, *UL29.2* or GFP siRNA and infected with HSV-2, peak *UL27* expression (6 h after infection) was significantly downregulated in response to *UL27.2*, but not to *UL29.2* or GFP siRNA ($P < 0.004$). Conversely, *UL29*, which is expressed earlier than *UL27*, was significantly downregulated both at 4 h and 6 h, and only in response to *UL29.2* siRNA ($P < 0.0001$ compared with GFP siRNA) (Fig. 2d). One day later, when infection had amplified by cell-to-cell spread, the expression of all four viral genes examined (siRNA-targeted *UL5*, *UL27* and *UL29* as well as the viral thymidine kinase *TK*) was reduced by siRNAs targeting any of the viral genes (Fig. 2e). These differences were all highly statistically significant. Even the least effective siRNA (*UL29.1*) reduced viral replication (that is, *TK* expression; $P < 0.002$ compared with GFP siRNA). Control GFP siRNA did not affect viral gene transcription. Viral gene silencing roughly paralleled the inhibition of viral replication, with *UL29.2* siRNA proving the most effective, suppressing relative viral gene expression by 4–5-fold ($P < 0.001$ compared with GFP siRNA). *UL5.2* and *UL27.2* siRNAs each inhibited viral gene expression by ~3-fold ($P < 0.002$ for *UL5.2*, $P < 0.001$ for *UL27.2* compared with GFP).

To investigate whether siRNAs could protect mice from HSV-2 infection, groups of 5–10 medroxyprogesterone-pretreated mice were given lipid-complexed *UL29.2* intravaginally 2 h before and 4 h after vaginal challenge with 2 LD₅₀ (2×10^4 plaque-forming units, p.f.u.) of HSV-2 wild-type virus. Mice treated with <250 pmol *UL29.2* siRNA were not protected, mice treated with 250 pmol siRNA were partially protected, and 500 pmol siRNA gave optimal protection (data not shown). We therefore administered 500 pmol siRNA in subsequent experiments.

UL29.2 siRNA provided highly significant protection, as assessed daily by a clinical disease scoring system or by survival (Fig. 3a, b). Although 75% of infected mice treated with GFP siRNA (15/20) or no siRNA (13/17) died, only 25% of mice treated with *UL29.2* (5/20) died (time to death comparison by log-rank test: $P < 0.001$ versus no treatment, $P < 0.003$ versus GFP siRNA). Although 55% of *UL29.2*-treated mice developed some signs of infection, surviving

¹CBR Institute for Biomedical Research, ²Department of Pediatrics, ³Department of Microbiology and Molecular Genetics, ⁴Dana Farber Cancer Institute and ⁵Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA.

mice were free of clinical disease by day 11. A longitudinal regression analysis of disease severity over time and between groups showed robust protection in UL29.2-treated mice ($P < 0.001$ versus no treatment, $P < 0.006$ versus GFP siRNA when analysed with respect to time course; $P < 0.001$ versus either control when analysed between groups).

Mice treated with UL27.2, which was less effective *in vitro*, were less effectively protected. Sixty per cent (6/10) of mice survived the lethal vaginal challenge ($P < 0.009$ compared with untreated, $P = 0.10$ compared with GFP siRNA). UL27.2 protection from disease severity was significant by longitudinal regression analysis ($P < 0.001$ compared with untreated, $P < 0.005$ compared with GFP siRNA with respect to time; $P < 0.01$ and $P = 0.05$ when analysed between the respective groups). The clinical advantage was also evident by quantifying shed virus six days after infection

(Fig. 3c). Although all infected mice not given siRNAs shed virus on day six, no virus was detected in 70% of UL29.2- and 50% of UL27.2-treated mice. No virus was isolated from three out of nine GFP siRNA-treated mice, but this was not significantly different from mice not treated with siRNAs. Comparison of virus recovered from UL29.2 siRNA-treated mice with GFP siRNA-treated mice also was not significant ($P = 0.09$ by Wilcoxon rank sum test). However, the geometric mean viral titre was reduced from $1,226 \text{ p.f.u. ml}^{-1}$ in untreated mice to $7.9 \text{ p.f.u. ml}^{-1}$ in mice that received UL29.2 ($P < 0.01$). Viral shedding at day six predicted survival, as 18 out of 19 mice from which virus was cultured died, whereas none out of 15 mice with undetectable virus died.

One concern about using RNAi against viruses is escape from RNAi by mutation of the targeted sequence. Escape mutation has been shown for polio, HIV and hepatitis C^{9–11}. We cloned and

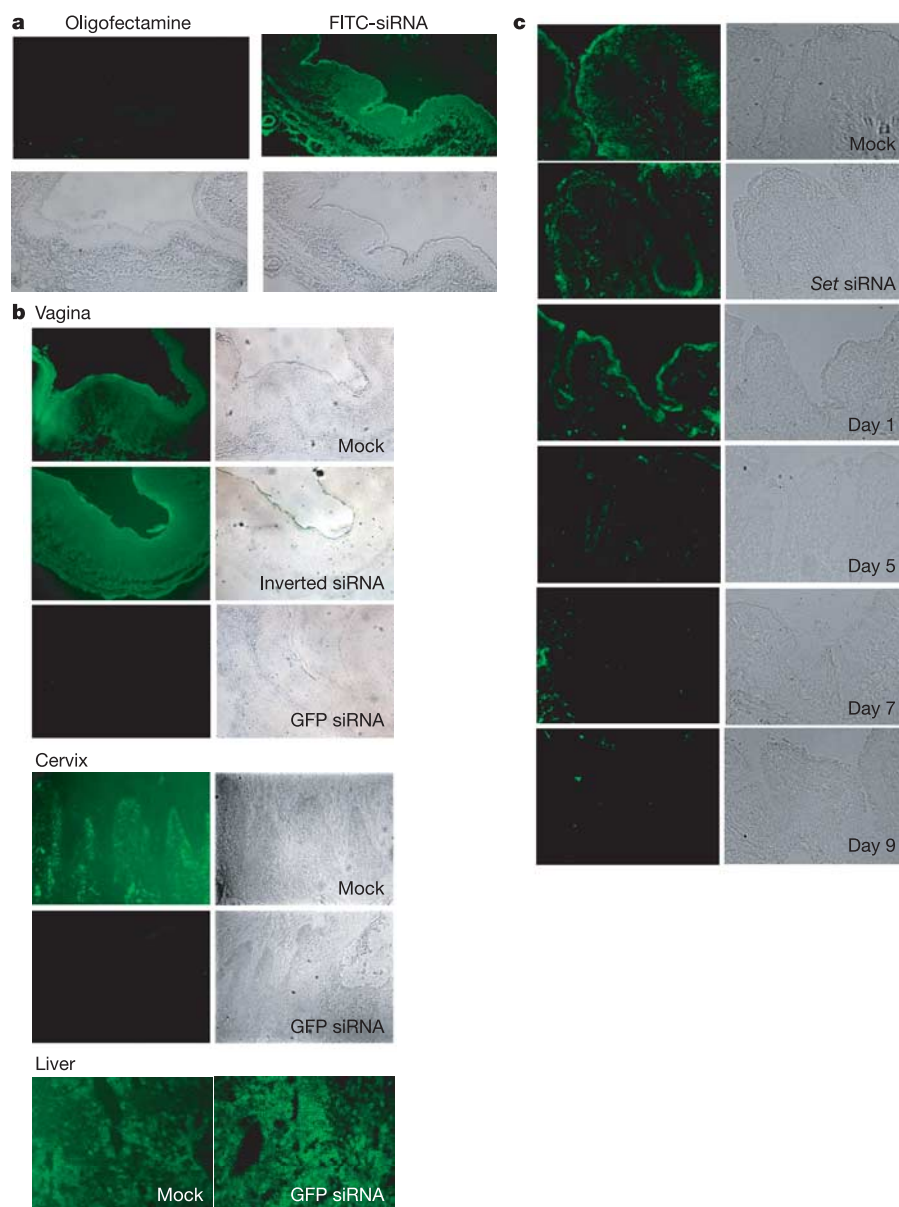


Figure 1 | siRNAs administered intravaginally are efficiently taken up by vaginal tissue and durably silence endogenous EGFP expression.

a, FITC-siRNA mixed with Oligofectamine is efficiently taken up throughout the mucosa and submucosa. Sections were obtained 24 h after administration. **b**, siRNA targeting *EGFP*, but not an inverted control sequence, silences *EGFP* expression throughout the mouse vagina and cervix

in GFP transgenic mice three days after administration. Liver *EGFP* expression is unaffected. **c**, Silencing persists for at least nine days in the vagina of GFP transgenic mice treated with GFP siRNAs. Data are representative of at least two experiments. An siRNA targeting an irrelevant gene (*Set*) was administered to control mice.

sequenced HSV-2 DNA from the day 6 vaginal swab from one UL29.2-treated mouse that died and from one control mouse. No mutations were found in 150-nucleotide stretches of *UL29*, which included the targeted sequence, in 24 sequences analysed from each mouse. Escape mutation is not anticipated to be as problematic for DNA viruses (such as HSV-2) as for RNA viruses.

The cervicovaginal mucosa of siRNA-treated, HSV-2-infected mice at day 6 was also spared (Fig. 3d). In control infected mice that were pretreated with no siRNA or GFP siRNA, the mucosal epithelium was partially denuded, and dying cells and inflammatory infiltrates were prominent. Multinucleated cells with intranuclear inclusion bodies—a hallmark of HSV-2 infection—were also evident. In contrast, in UL27.2 or UL29.2 siRNA-treated mice, the epithelium

was intact and there were few apoptotic bodies and scarcely any inflammatory cells.

To investigate the effects of delaying siRNA treatment until after HSV-2 exposure, 500 pmol of UL27.2 or UL29.2, or a mixture of both (250 pmol each), was administered intravaginally 3 and 6 h after infection. Mice receiving UL27.2 or UL29.2 alone had no survival advantage compared with mice given GFP siRNA (2/6 survived) or no siRNA (1/6 survived) (Fig. 3e). However, 5/6 mice given both UL27.2 and UL29.2 siRNA survived ($P = 0.11$ compared with GFP siRNA; $P < 0.04$ compared with no siRNA). Therefore, post-exposure treatment might be effective. Targeting multiple genes will probably work better than targeting a single gene.

Under certain circumstances, siRNAs can induce the interferon (IFN) pathway and trigger inflammation^{12–15}. We therefore analysed vaginal tissue for inflammatory infiltrates (Fig. 4a) and induction of interferon and interferon-responsive genes 24 and 48 h after siRNA treatment (Fig. 4b). siRNA treatment did not cause an inflammatory infiltrate. Moreover, *Ifnb* and the principal interferon-responsive genes, *Oas1* and *Stat1*, were not significantly induced when analysed by quantitative RT-PCR. As expected, HSV-2 infection in the absence of siRNAs (used as a positive control) activated interferon-responsive genes.

Vaginal instillation of siRNAs targeting essential viral genes protects mice from vaginal challenge with a lethal dose of HSV-2. The treatment was well-tolerated without causing inflammation or inducing interferon-responsive genes. This efficient and lasting silencing deep in the vaginal tissue was unexpected, and augurs well for using siRNAs to prevent or treat sexually transmitted viral and parasitic infections. Our results, together with impressive results in lung models of viral infection^{16–20}, suggest that siRNA uptake at mucosal surfaces may be particularly efficient and involve mechanisms not present in internal organs.

Much work needs to be done to develop siRNAs as the basis for a microbicide. These experiments were done without optimizing the siRNAs for silencing efficiency or chemical modifications that enhance resistance to endogenous RNases²¹. siRNAs would also need to be formulated in a vehicle acceptable for vaginal retention. The effect of menstrual variation on protection, especially on the durability of silencing, needs to be evaluated. Viral sequence variability also needs to be addressed. However, by targeting relatively well-conserved viral sequences in essential viral genes or by combining siRNAs that target multiple viral genes, the related problems of viral sequence diversity and potential escape mutation might be mitigated. Although we did not find evidence of escape mutation, this might take longer than six days to develop. Any extension of our results to the designing of an HIV microbicide would also require demonstrating silencing in resident tissue macrophages, dendritic cells and T cells, which are rare in normal, uninfamed vaginal tissue.

Finally, cost is an important consideration for a microbicide designed for global use. Only 500 pmol siRNA was required to protect mice in this study. The manufacturing cost of a single application for humans, crudely estimated on the basis of scaling up by weight and current costs, is \$8. If silencing is durable and treatments can be spaced, this is a realistic cost. Given the devastating global epidemic and the unlikelihood of there being an effective HIV-1 vaccine soon, we feel that investigating whether RNAi can be harnessed for use in microbicides is a sensible approach.

Note added in proof: In the advance online publication of this Letter, in the second sentence of the fourth paragraph '25 μ M, and reached a plateau at 100 μ M' should read '25 nM, and reached a plateau at 100 nM'. In addition, the x axis of Fig. 2c should read 'siRNA concentration (nM)'. These errors have been corrected for print.

METHODS

Mice. BALB/c mice (5–8 weeks old) were obtained from Taconic Farms; FVB.Cg-Tg(GFP)5Nagy mice were from Jackson Laboratories⁷. Mice were

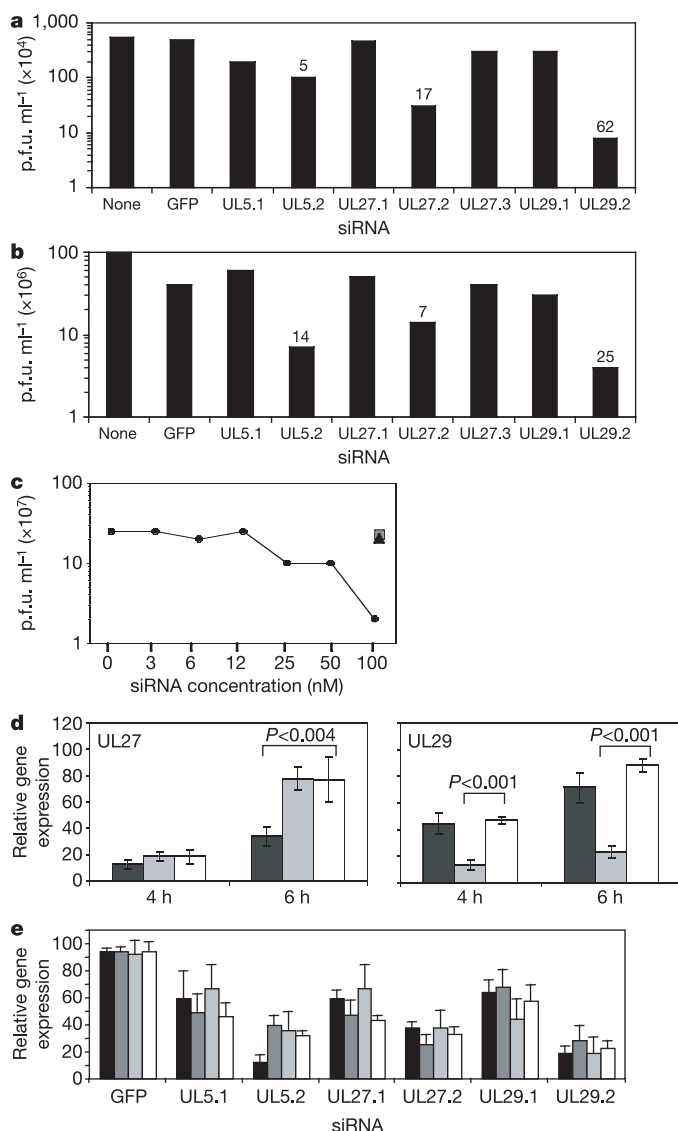


Figure 2 | siRNAs targeting HSV-2 reduce viral replication. NIH3T3 (a) or Vero (b–e) cells were transfected overnight with siRNA, then infected with HSV-2 and harvested 20 h later. Values above bars show fold reduction in viral plaques. Data are representative of five independent experiments. c, Dose-response curve, showing effect of treatment with UL29.2 (filled circles), inverted UL29.2 (filled square) or GFP (filled triangle) siRNA. d, e, Gene silencing by real-time RT-PCR was specific at 4 h or 6 h, that is, before cell-to-cell spread (UL27.2, dark grey; UL29.2, light grey; GFP, white) (d), but expression of all viral genes (*UL5*, black; *UL27*, dark grey; *UL29*, light grey; *TK*, white) was suppressed at 24 h (e). Data show mean $\pm$ s.d. from one of two experiments.

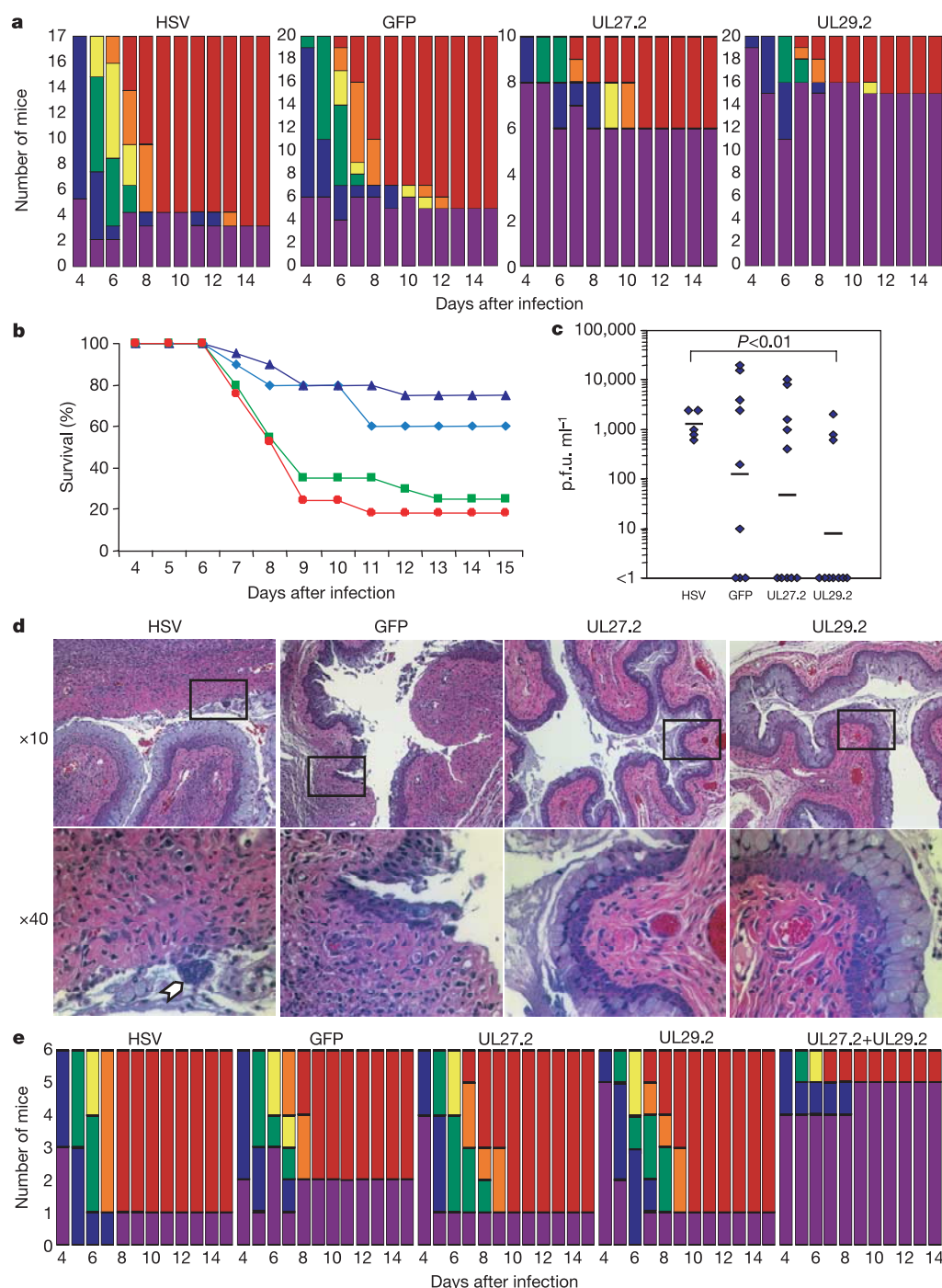


Figure 3 | siRNAs protect mice from lethal HSV-2 infection. **a–d**, Mice given lipid-complexed siRNA intravaginally 2 h before and 4 h after infection with ~ 2 LD₅₀ HSV-2 were analysed for disease severity (**a**) (see colour code provided in Methods), survival (**b**) (HSV only, red; GFP, green; UL27.2, light blue; UL29.2, dark blue), viral shedding on day 6 (**c**) and cervicovaginal histopathology on day 6 (**d**). **a**, **b** show data from three experiments. Transfection of lipid alone did not affect HSV-2 disease (not shown).

c, Vaginal viral shedding. Bars represent geometric mean titre. **d**, The epithelium is preserved after UL27.2 or UL29.2 siRNA treatment, with decreased inflammatory infiltrates and fewer dying cells. Boxes indicate areas magnified in lower panels. White arrow points to a multinucleated cell with viral inclusion—a hallmark of HSV-2 infection. **e**, A combination of UL27.2 plus UL29.2 siRNA, but neither siRNA alone, protects from HSV-2 disease after exposure. Data are representative of two experiments.

subcutaneously injected with 2 mg medroxyprogesterone acetate (Sicor), and then 1 week later were infected vaginally with 2×10^4 p.f.u. (~ 2 LD₅₀) HSV-2 strain 186 (ref. 22). siRNA (500 pmol) was complexed with Oligofectamine (Invitrogen) according to the manufacturer's protocol, and was then administered intravaginally (in a maximum volume of 12 μ l) either 2 h before and 4 h after HSV-2 infection or 3 h and 6 h after HSV-2 infection. Clinical signs of infection were graded according to a five-point scale: 0, no signs of infection (purple); 1, slight genital erythema and oedema (blue); 2, moderate genital inflammation (green); 3, purulent genital lesions (yellow); 4, hind limb paralysis

(orange); 5, death (red)²². Viral shedding was determined by swabbing the vaginal cavity (using a Micropur swab, PurFybr Inc.) on day 6 after infection, and titrating the virus on Vero cells. In some cases, the vagina was dissected at the indicated times and either fixed in 10% formalin (Sigma) for paraffin embedding and sectioning, or stored in RNAlater (Qiagen) for RNA isolation.

Viruses and transfection assays. For *in vitro* studies, 186 Δ Kpn, a replication-competent, TK-negative mutant of strain 186syn + (ref. 23) was grown in Vero cells as described²⁴. Vero or NIH3T3 cells (ATCC) (4×10^5 cells per well in 6-well plates in 1 ml of complete medium, plated one day earlier), were treated

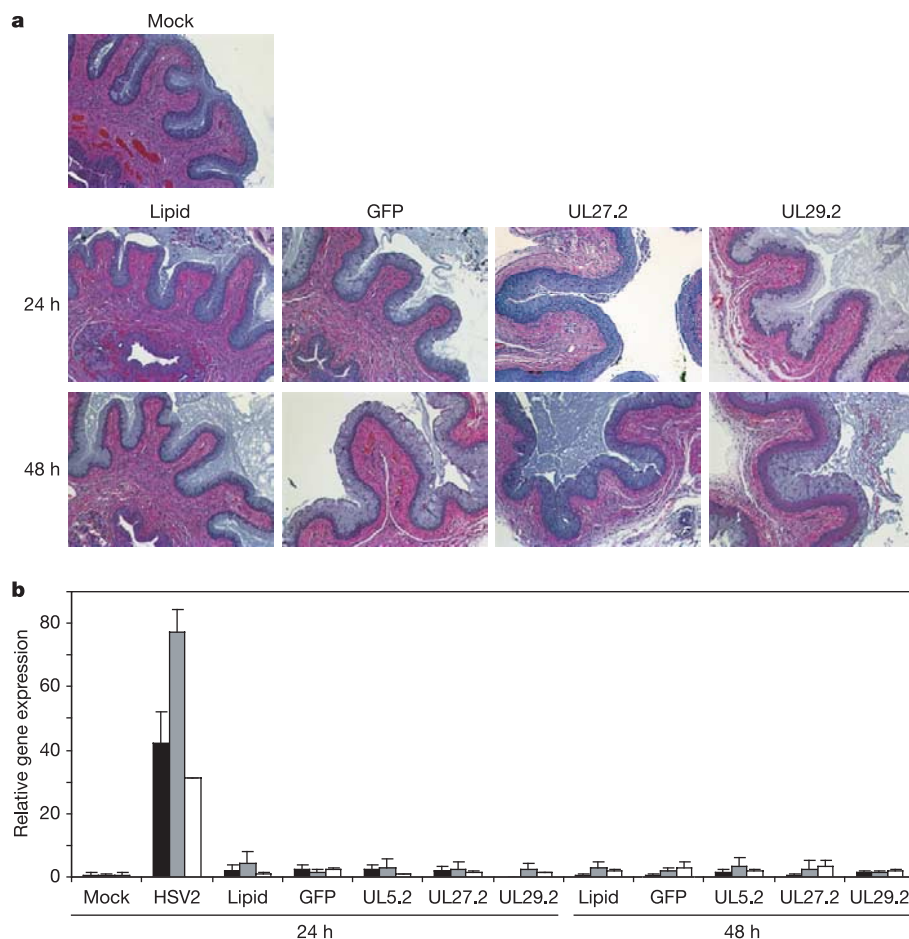


Figure 4 | Topical lipid-complexed siRNAs do not activate inflammation or interferon-responsive genes. **a**, **b**, Vaginal tissue, dissected 24 h or 48 h after administering 500 pmol of lipid-complexed siRNA, was assessed by haematoxylin-eosin staining for inflammation (**a**, $\times 10$ magnification) and by quantitative RT-PCR for expression of *Ifnb* (black) and the interferon-responsive genes *Stat1* (grey) and *Oas1* (white) relative to the control gene

Gapdh (**b**). HSV-2 infection was used as a positive control for interferon induction. In siRNA-treated mice, no HSV-2 was administered. None of the siRNAs induced a significant change in interferon-responsive gene expression compared to mock-treated mice given only PBS intravaginally. Data show mean $\pm$ s.d.

with 100 pmol or the indicated concentration of siRNA. The siRNA had been complexed with TransIT-TKO (Mirus) to transfect Vero cells or with TransIT-siQuest (Mirus) for NIH3T3 cells, according to the manufacturer's instructions. The medium was replaced after overnight incubation at 37 °C, and 2 h later HSV-2 186 Δ Kpn was added at an MOI of 1. After 1 h at 37 °C, the medium was again replaced. Cells were harvested 24 h later and viral titre determined by plaque assay on Vero cells. For mouse experiments, wild-type HSV-2 strain 186syn + virus was used²⁵. An aliquot of virus used for each mouse experiment was also assayed by plaque assay to confirm viral titre.

siRNAs. siRNAs (Dharmacon) were prepared according to the manufacturer's instructions. FITC-labelled siRNA was a previously described sequence targeting CD4 (ref. 26). The sequence for silencing *EGFP* has been described²⁶. The sequences for HSV-2 (GenBank accession number NC 001798) siRNAs were: UL5.1 (nt 12838–12856) sense 5'-CUACGGCAUCAGCUCCAAA-3', antisense 5'-UUUGGAGCUGAUGCCGUAG-3'; UL5.2 (nt 12604–12622) sense 5'-UGUGGUCAUUGUCUAUUA-3', antisense 5'-UUAUAGACAAUGACCACA-3'; UL27.1 (nt 54588–54606) sense 5'-GUUUACGUUAUACCACAU-3', antisense 5'-UAUGUGGUUAUACGUAAAC-3'; UL27.2 (nt 54370–54388) sense 5'-ACGUGAUCGUGCAGAACUC-3', antisense 5'-GAGUUCUGCAGCAUCA-3'; UL27.3 (nt 54097–54115) sense 5'-UCGACCUGAACAUACCAU-3', antisense 5'-AUGGUGAUGUUCAGGUCGA-3'; UL29.1 (nt 59715–59733) sense 5'-CCACUCGACGUACUUAUA-3', antisense 5'-UAUGAAGUACGUCGAGUGG-3'; UL29.2 (nt 60324–60342) sense 5'-CUUUCGCAUCAAUUC-3', antisense 5'-UUGGAAUUGAUUGCGAAAG-3'; inverted UL29.2 sense 5'-AACCUAAACUACGCUUUC-3', antisense 5'-GAAAGCGUUAAGGUU-3'.

Quantitative RT-PCR. Total RNA (1 μ g) was isolated using the RNeasy RNA isolation kit (Qiagen) and reverse transcribed using Superscript III (Invitrogen)

and random hexamers, according to the manufacturer's protocol. Real-time PCR was performed on 0.2 μ l of complementary DNA, or a comparable amount of RNA with no reverse transcriptase, using Platinum Taq Polymerase (Invitrogen) and a Biorad iCycler. SYBR green (Molecular Probes) was used to detect PCR products. Reactions were performed in 25 μ l in triplicate. Primers were: *Gapdh* forward 5'-TTCACCACCATGGAGAAGGC-3', *Gapdh* reverse 5'-GGCATGGACTGTGGTCATGA-3', *TK* forward 5'-CGATCTACTCGCCAA CACGGTG-3', *TK* reverse 5'-GAACGCGGAACAGGGCAAACAG-3', *UL5* forward 5'-TCGCTGGAGTCCACCTTCGAAC-3', *UL5* reverse 5'-CGAACTC GTGCTCCACACATCG-3', *UL27* forward 5'-CAAAGACGTGACCGTGTG CAG-3', *UL27* reverse 5'-GCGGTGGTCTCCATGTTGTTCC-3', *UL29* forward 5'-GCCAGGAGATGGACGTGTTTCG-3', *UL29* reverse 5'-CGCGCTGTT CATCGTTCCGAAG-3', *Stat1* forward 5'-TTTGCCAGACTCGAGCTCCTG-3', *Stat1* reverse 5'-GGGTGCAGGTTCCGGATTCAAC-3', *Oas1* forward 5'-GGAGGTTGCAGTGCCCAACGAAG-3', *Oas1* reverse 5'-TGGAAGGGAGGCA GGGCATAAC-3', *Ifnb* forward 5'-CTGGAGCAGCTGAATGAAAG-3', *Ifnb* reverse 5'-CTTGAAGTCCGCCCTGTAGGT-3'.

PCR parameters consisted of 5 min Taq activation at 95 °C, followed by 40 cycles of 95 °C $\times$ 20 s, 60 °C $\times$ 30 s, and 69 °C $\times$ 20 s. Standard curves were generated and the relative amount of mRNA was normalized to *Gapdh* mRNA. Specificity was verified by melt curve analysis and agarose gel electrophoresis.

Tissue sections and microscopy. For fluorescence microscopy, dissected tissue was placed in optimal cutting temperature compound (TissueTek) and snap-frozen in LN₂. For haematoxylin-eosin stained sections, tissues were fixed in 10% formalin and paraffin-embedded. Microscopy was performed and scored (by an operator blind to the treatment condition) on a Zeiss Axiovert 200M microscope using Slidebook acquisition and analysis software (Intelligent Imaging).

Statistical analysis. *In vitro* data were analysed by Student's *t*-test. Survival distribution was calculated using the Kaplan and Meier method²⁷, and the univariate comparison of survival for control versus treated groups was tested using a log-rank test, comparing two groups at a time²⁸. The approach of generalized estimating equations was used to model disease scores collected over time and to compare disease severity of control versus treated groups²⁹. All *P*-values are for two-tailed significance tests.

Received 12 July; accepted 26 September 2005.

Published online 23 November 2005.

- Whitley, R. J. in *Field's Virology* (eds Knipe, D. M. & Howley, P. M.) 2461–2510 (Lippincott, Williams and Wilkins, Philadelphia, 2001).
- Wald, A. & Link, K. Risk of human immunodeficiency virus infection in herpes simplex virus type 2-seropositive persons: a meta-analysis. *J. Infect. Dis.* **185**, 45–52 (2002).
- Celum, C., Levine, R., Weaver, M. & Wald, A. Genital herpes and human immunodeficiency virus: double trouble. *Bull. World Health Organ.* **82**, 447–453 (2004).
- Pilcher, H. Starting to gel. *Nature* **430**, 138–140 (2004).
- Shankar, P., Manjunath, N. & Lieberman, J. The prospect of silencing disease using RNA interference. *J. Am. Med. Assoc.* **293**, 1367–1373 (2005).
- Roizman, B. & Knipe, D. M. in *Field's Virology* (eds Knipe, D. M. & Howley, P. M.) 2399–2440 (Lippincott, Williams and Wilkins, Philadelphia, 2001).
- Hadjantonakis, A. K., Gertsenstein, M., Ikawa, M., Okabe, M. & Nagy, A. Generating green fluorescent mice by germline transmission of green fluorescent ES cells. *Mech. Dev.* **76**, 79–90 (1998).
- Khvorov, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).
- Boden, D., Pusch, O., Lee, F., Tucker, L. & Ramratnam, B. Human immunodeficiency virus type 1 escape from RNA interference. *J. Virol.* **77**, 11531–11535 (2003).
- Gitlin, L., Stone, J. K. & Andino, R. Poliovirus escape from RNA interference: short interfering RNA-target recognition and implications for therapeutic approaches. *J. Virol.* **79**, 1027–1035 (2005).
- Wilson, J. A. & Richardson, C. D. Hepatitis C virus replicons escape RNA interference induced by a short interfering RNA directed against the NS5b coding region. *J. Virol.* **79**, 7050–7058 (2005).
- Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H. & Williams, B. R. Activation of the interferon system by short-interfering RNAs. *Nature Cell Biol.* **5**, 834–839 (2003).
- Heidel, J. D., Hu, S., Liu, X. F., Triche, T. J. & Davis, M. E. Lack of interferon response in animals to naked siRNAs. *Nature Biotechnol.* **22**, 1579–1582 (2004).
- Hornung, V. *et al.* Sequence-specific potent induction of IFN- α by short interfering RNA in plasmacytoid dendritic cells through TLR7. *Nature Med.* **11**, 263–270 (2005).
- Judge, A. D. *et al.* Sequence-dependent stimulation of the mammalian innate immune response by synthetic siRNA. *Nature Biotechnol.* **23**, 457–462 (2005).
- Ge, Q. *et al.* Inhibition of influenza virus production in virus-infected mice by RNA interference. *Proc. Natl Acad. Sci. USA* **101**, 8676–8681 (2004).
- Tompkins, S. M., Lo, C. Y., Tumpey, T. M. & Epstein, S. L. Protection against lethal influenza virus challenge by RNA interference *in vivo*. *Proc. Natl Acad. Sci. USA* **101**, 8682–8686 (2004).
- Bitko, V., Musiyenko, A., Shulyayeva, O. & Barik, S. Inhibition of respiratory viruses by nasally administered siRNA. *Nature Med.* **11**, 50–55 (2005).
- Zhang, W. *et al.* Inhibition of respiratory syncytial virus infection with intranasal siRNA nanoparticles targeting the viral NS1 gene. *Nature Med.* **11**, 56–62 (2005).
- Li, B. J. *et al.* Using siRNA in prophylactic and therapeutic regimens against SARS coronavirus in Rhesus macaque. *Nature Med.* **11**, 944–951 (2005).
- Manoharan, M. RNA interference and chemically modified small interfering RNAs. *Curr. Opin. Chem. Biol.* **8**, 570–579 (2004).
- Morrison, L. A., Da Costa, X. J. & Knipe, D. M. Influence of mucosal and parenteral immunization with a replication-defective mutant of HSV-2 on immune responses and protection from genital challenge. *Virology* **243**, 178–187 (1998).
- Jones, C. A., Taylor, T. J. & Knipe, D. M. Biological properties of herpes simplex virus 2 replication-defective mutant strains in a murine nasal infection model. *Virology* **278**, 137–150 (2000).
- Gao, M. & Knipe, D. M. Genetic evidence for multiple nuclear functions of the herpes simplex virus ICP8 DNA-binding protein. *J. Virol.* **63**, 5258–5267 (1989).
- Spang, A. E., Godowski, P. J. & Knipe, D. M. Characterization of herpes simplex virus 2 temperature-sensitive mutants whose lesions map in or near the coding sequences for the major DNA-binding protein. *J. Virol.* **45**, 332–342 (1983).
- Novina, C. D. *et al.* siRNA-directed inhibition of HIV-1 infection. *Nature Med.* **8**, 681–686 (2002).
- Kaplan, E. L. & Meier, R. Non-parametric estimation from incomplete observation. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
- Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–170 (1966).
- Zeger, S. L. & Liang, K. Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130 (1986).

Acknowledgements We thank R. Colgrove, T. Taylor, E. Torres-Lopez, D. Brown and S. White for advice. This work was supported by grants from the NIH to D.M.K. and J.L., and by postdoctoral fellowships from the Harvard Center for AIDS Research and amfAR to D.P. and the Leukemia and Lymphoma Society to D.C.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.L. (lieberman@cbr.med.harvard.edu).

Potential of neuroblastoma metastasis by loss of caspase-8

Dwayne G. Stupack^{1*}, Tal Teitz^{2*}, Matthew D. Potter¹, David Mikolon¹, Peter J. Houghton³, Vincent J. Kidd², Jill M. Lahti² & David A. Cheresch¹

Neuroblastoma, the most common paediatric solid tumour, arises from defective neural crest cells¹. Genetic alterations occur frequently in the most aggressive neuroblastomas¹. In particular, deletion or suppression of the proapoptotic enzyme caspase-8 is common in malignant, disseminated disease, although the effect of this loss on disease progression is unclear^{2–4}. Here we show that suppression of caspase-8 expression occurs during the establishment of neuroblastoma metastases *in vivo*, and that reconstitution of caspase-8 expression in deficient neuroblastoma cells suppressed their metastases. Caspase-8 status was not a predictor of primary tumour growth; rather, caspase-8 selectively potentiated apoptosis in neuroblastoma cells invading the collagenous stroma at the tumour margin. Apoptosis was initiated by unligated integrins by means of a process known as integrin-mediated death⁵. Loss of caspase-8 or integrin rendered these cells refractory to integrin-mediated death, allowed cellular survival in the stromal microenvironment, and promoted metastases. These findings define caspase-8 as a metastasis suppressor gene that, together with integrins, regulates the survival and invasive capacity of neuroblastoma cells.

To investigate the impact of caspase-8 loss on neuroblastoma dissemination, we tested clinically derived neuroblastoma lines^{3,4} for their capacity to metastasize in chick embryos^{6,7}. This model permits the assessment of tumour growth, cell invasion and spontaneous metastasis to sites such as bone marrow, within the same animal. These events do not readily occur in adult mice⁸. Although caspase-8 expression did not predict primary tumour growth, we observed significantly more apoptosis (detected by staining using TdT-mediated dUTP nick end labelling (TUNEL) and by digital analysis of colocalization) in locally invasive caspase-8-positive extratumoral cells relative to the caspase-8-negative cells (Fig. 1 and Supplementary Fig. S1a), indicating that caspase-8 expression sensitized the stroma-invasive neuroblastoma cells to apoptosis. Similarly, reconstitution of caspase-8-deficient cells with physiological levels of wild-type caspase-8 (Supplementary Fig. S1c), but not an inactive mutant, yielded increased apoptosis in tissue-invasive cells (Fig. 1). Apoptosis levels were similar within the primary tumour mass regardless of caspase-8 expression (Fig. 1b), indicating that apoptosis might have been selective for tissue-invasive cells.

We assessed whether caspase-8 expression influenced the spontaneous metastases of neuroblastoma to bone or lung tissue, unfavourable events in human disease⁹. Although primary neuroblastoma tumours grew to about the same size, irrespective of caspase-8 expression (Fig. 2a, left panel), lung and bone marrow metastases were primarily detected in embryos bearing tumours deficient in caspase-8, with an average incidence of about 35%

(Fig. 2a, right panels). Metastases were observed rarely in embryos bearing caspase-8-positive tumours (about 7%), with the exception of NB16 cells, which were detected in bone marrow with an incidence of about 17% (Fig. 2a, right panels). Reconstitution of caspase-8 expression in caspase-8-negative NB7 or NB10 cells did not impact primary tumour growth but significantly suppressed metastasis (Fig. 2b), whereas an inactive point mutant of caspase-8 (C8₃₆₀) was unable to prevent metastasis (Fig. 2b). Knockdown of endogenous (NB5, NB16, SKNAS) or ectopic (NB7C8) caspase-8 expression by RNA-mediated interference promoted tumour cell dissemination to lung and bone marrow but did not influence primary tumour growth (Fig. 2c). These findings were reflected in a decreased incidence of apoptosis in the caspase-8 knockdown cells invading the proximal chorioallantoic membrane (CAM) tissue (Supplementary Fig. S1). Thus, caspase-8 is an important determinant in cell survival during metastasis but not primary tumour growth.

We next examined whether caspase-8 suppression was associated with spontaneous metastases from syngeneic murine neuroblastoma. The non-metastatic murine neuroblastoma cell line C1300 forms sized primary tumours similar in size to those in two metastatic sublines, NXS2 and TBJ^{10,11}. Both metastatic sublines were deficient in caspase-8, whereas the non-metastatic parental cells retained caspase-8 expression (Fig. 3a). Caspase-9 and caspase-10 expression were maintained in both the parental and metastatic tumour lines (Fig. 3a), which is consistent with results in human disease. As in mice^{10,11}, the NXS2 and TBJ cells were highly metastatic in the chick embryo (Supplementary Fig. S2), whereas parental C1300 cells metastasized poorly, indicating that loss of caspase-8 might facilitate metastasis in distinct animal models of this disease.

Although the human neuroblastomata are inefficient at forming spontaneous metastases in mice⁸, we were able to generate experimental ovarian, adrenal, renal and hepatic metastases 9–12 weeks after intravenous injection of tumour cells. These metastases showed decreased transcription of the transgenic caspase-8 mRNA, as well as that of a green fluorescent protein (GFP) marker translated from the same message by means of an internal ribosome-binding sequence, although control cells containing the empty retroviral vector retained GFP expression. Caspase-8 mRNA levels were decreased in metastases relative to input cells (Fig. 3b) or non-metastatic neuroblastoma tumours grown subcutaneously (Supplementary Fig. S3). Immunoblotting confirmed the loss of caspase-8 protein expression in both metastatic lesions and cell lines derived from these lesions (Fig. 3c). These data support the notion that loss of caspase-8 expression promotes metastatic disease^{3,4}.

Caspase-8 mediates apoptosis initiated by death receptors such as Fas^{12,13}. Whereas cycloheximide presensitizes neuroblastoma to

¹Department of Pathology and The John and Rebecca Moores Cancer Center, The University of California at San Diego, La Jolla, California 92093-0803, USA. ²Department of Genetics and Tumor Cell Biology, and ³Department of Molecular Pharmacology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA.

*These authors contributed equally to this work.

death-receptor-mediated apoptosis^{3,4}, the apoptosis observed *in vivo* (Fig. 1) occurred in the absence of cycloheximide, where neuroblastomas are generally resistant to death-receptor-mediated killing (Supplementary Fig. S4a–c). Caspase-8 expression does not predict susceptibility to proapoptotic agents such as etoposide or staurosporine, which act through caspase-9 (ref. 14; Supplementary Fig. S4d–f). Caspase-8-expressing neuroblastomas are therefore not

generally ‘apoptosis-prone’, but rather undergo specific caspase-8-dependent apoptosis during tissue invasion in a death-receptor-independent manner.

Caspase-8 activation can occur in a manner independent of

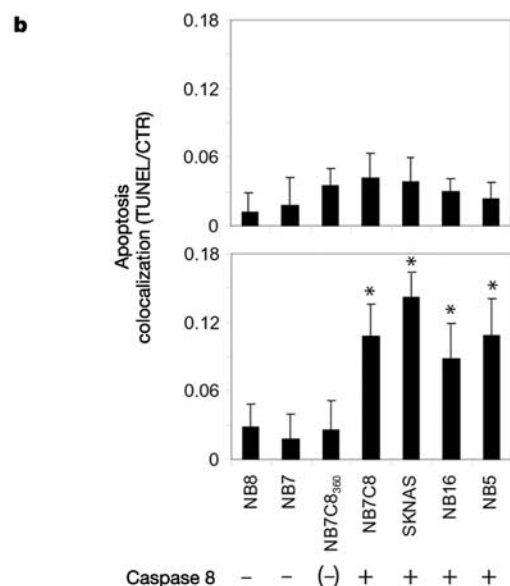
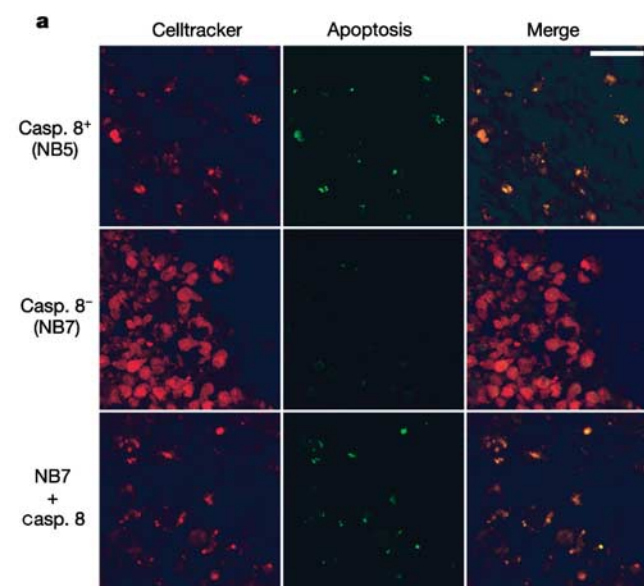


Figure 1 | Increased survival among caspase-8-deficient neuroblastoma invading stroma. **a**, Celltracker-labelled (red) neuroblastoma cells expressing endogenous caspase-8 (casp. 8⁺, NB5 neuroblastoma cells), deficient in caspase-8 (casp. 8⁻, NB7 neuroblastoma cells) or reconstituted with caspase-8 (NB7 + casp. 8) were seeded in the chorioallantoic membrane. Individual cells were observed invading the stroma along the tumour margin. Apoptosis was evaluated by TUNEL staining (FITC, green). Scale bar, 100 μ M. **b**, Digital analysis of colocalization was performed³. Upper panel, tumour; lower panel, stroma. Five images were collected in the margin (1–2 mm from the tumour parenchyma) of each tumour, with a total of six to eight tumours assessed for each cell line. Data are mean $\pm$ s.e.m. of the incidence of green/red pixel colocalization. Asterisk, $P < 0.05$ for increased TUNEL staining among the caspase-8-positive cell lines relative to the reference line NB7.

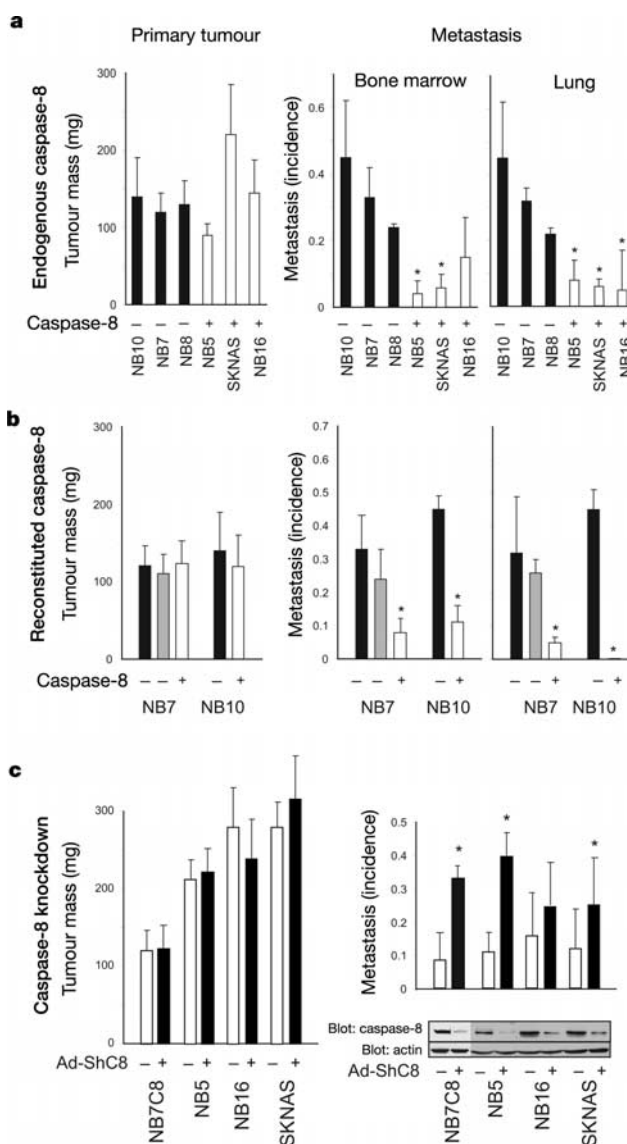


Figure 2 | Assessment of neuroblastoma tumour formation and spontaneous metastasis in the chick embryo. Left panels show primary tumour growth; right panels represent metastasis to bone marrow and lung, respectively. **a**, Primary neuroblastoma tumours derived from cells expressing (white bars) or lacking (black bars) caspase-8 expression were resected and then weighed. The incidence of metastasis to bone marrow or lung was determined for each neuroblastoma tumour line by PCR analysis of total tissue DNA for the presence of human Alu sequence⁷. **b**, Primary tumour formation and metastasis were assessed in caspase-8-deficient neuroblastoma cells reconstituted with caspase-8. Parental tumours (black bar) or sublines reconstituted with physiological levels of active (white bars) or inactive caspase-8 (C8360, grey bars) were assessed as described above. **c**, Knockdown of endogenous (NB5, NB16, SKNAS) or ectopic (NB7C8) caspase-8 expression by shRNA-C8 encoding adenovirus (+) was performed, and the impact on tumour growth and metastasis to either site was determined as described above. Caspase-8 expression was assessed by immunoblot analysis relative to empty vector virus (–) (inset). Data are means $\pm$ s.e.m. The mean metastasis incidence observed in three to six separate experiments ($n = 8–20$ per experiment) is shown; χ^2 analysis was performed on the total cohort. A collective analysis of all data found the expression of caspase-8 to be a strong predictor of decreased metastasis ($P < 0.001$).

Fas-associated death domain protein (FADD) in cells present within an inappropriate extracellular matrix *in vitro* as part of an apoptotic response initiated by unligated or antagonized integrins⁵. The process, termed integrin-mediated death (IMD), occurs when caspase-8 becomes recruited to, and activated by, clusters of unligated or antagonized integrins at the cell surface^{5,15}. Thus, cells adherent within a three-dimensional collagen matrix that express high levels of non-collagen-binding integrins are induced to undergo apoptosis in a caspase-8-dependent manner. Interestingly, type I collagen is the principal matrix component of numerous stromal tissues, including the chorioallantois^{16,17}, and is not a ligand for integrins $\alpha_3\beta_1$, $\alpha_5\beta_1$ or $\alpha_v\beta_5$ expressed on the neuroblastoma cells used here (Supplementary Table 1).

Consistent with IMD was our observation that clustering and immunoprecipitation of unligated β_1 integrins from neuroblastoma cells resulted in the co-precipitation of caspase-8 (Supplementary Fig. S5). Accordingly, neuroblastoma cells expressing caspase-8 died rapidly after being seeded within a type I collagen matrix, whereas caspase-8-deficient cells remained viable and formed tumour colonies within this matrix (Fig. 4a). Neuroblastoma cells expressing an inactive caspase-8 mutant or caspase-8-positive neuroblastoma treated with the caspase inhibitor zVAD demonstrated increased survival in collagen (Fig. 4b). Similarly, neuroblastoma cells subjected to short hairpin RNA (shRNA)-mediated knockdown of caspase-8 showed increased survival (Fig. 4c, Supplementary Fig. S6). In contrast, expression of dominant-negative FADD, which blocks death-receptor-mediated killing, did not influence survival in three-dimensional collagen (Fig. 4c) or metastasis *in vivo* (Supplementary Fig. S7). Consistent with IMD was our observation that apoptosis of neuroblastoma cells within a collagen matrix was dependent on caspase-8 but independent of death receptors⁵.

Unligated integrins promote apoptosis by means of IMD in a manner similar to that of dependence receptors¹⁸. Accordingly, decreased expression of unligated integrins enhances cell survival in a collagen matrix⁵. The most abundant integrin on the matched NB7 and NB7C8 sublines was $\alpha_3\beta_1$, a laminin-binding integrin (Supplementary Table 1). To address the role of this integrin on NB7 and NB7C8 cell invasion and survival, we isolated subpopulations of these cells deficient in $\alpha_3\beta_1$ expression (Fig. 4d). The α_3 -deficient NB7C8 cells (NB7C8- α_3 L) showed increased survival relative to $\alpha_3\beta_1$ -expressing parental cells when seeded within collagen (Fig. 4e), and increased metastasis *in vivo* (Fig. 4f). Loss of integrin $\alpha_3\beta_1$ did not affect the survival or metastasis of caspase-8-deficient

NB7 cells (Fig. 4e, f). Antagonistic anti-integrin α_3 antibodies selectively enhanced the apoptosis of NB7C8 cells embedded in collagen and prevented their metastasis *in vivo* (Supplementary Fig. S4). Antibodies directed against α_v or α_5 integrin, which are expressed at much lower levels than $\alpha_3\beta_1$, did not significantly affect apoptosis or metastasis (Supplementary Fig. S5). Notably, the expression of $\alpha_3\beta_1$ was decreased both on the NB7C8 metastases isolated from severe combined immunodeficiency (SCID) mice and in clinical samples from patients with advanced stages of neuroblastoma (Supplementary Table S2). These data support the notion that resistance to IMD promotes the survival and dissemination of neuroblastoma cells, and the loss of either caspase-8 or integrins such as $\alpha_3\beta_1$ can potentiate metastatic disease.

IMD represents a new role for integrins in the regulation of cell invasion. In the unligated or antagonized state, integrins activate a caspase-8-dependent checkpoint and block cell invasion into an inappropriate microenvironment⁵. Our findings reveal that one mechanism to escape IMD occurs through the loss of caspase-8, resulting in enhanced survival during cell invasion (Supplementary Fig. S8). Restoration of caspase-8 sensitizes these cells to IMD, thereby compromising their invasive and metastatic properties. IMD does not seem to have a function in promoting apoptosis within the primary tumour, indicating that matrix remodelling within the tumour¹⁹, cell-cell interactions^{20,21} or autocrine factors²² might promote cell survival in this environment. IMD occurs among matrix-attached cells, differentiating this pathway from anoikis, or complete loss of integrin mediated attachment²³. The distinction between anoikis²⁴ and IMD⁵ is supported by the observation that decreased expression of integrins selectively enhances neuroblastoma survival and metastases. Ligated integrins mediate cell survival and the invasion of many cell types²⁵; the balance between the prosurvival and proapoptotic (IMD) integrin activities may therefore have a profound influence on the metastatic potential of a tumour cell.

Caspase-8 expression is lost in tumours other than neuroblastoma, including small-cell lung carcinoma²⁶, medulloblastoma²⁷ and about 10% of colorectal carcinomas²⁸. Whereas the loss or mutation of tumour suppressor genes such as p16 or p53 (ref. 29) enhances tumour growth, the loss of caspase-8 in neuroblastoma, and possibly other cancers, potentiates tumour cell invasion without affecting the growth of the primary tumour. These findings indicate that caspase-8 might be a metastasis suppressor gene³⁰ and that overcoming IMD by the suppression of caspase-8 activity, or through changes in integrin expression, facilitates tumour invasion by enhancing cell survival.

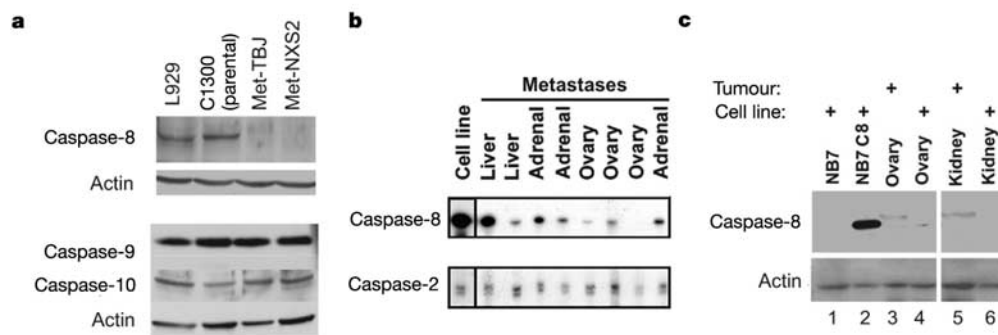


Figure 3 | Loss of caspase-8 occurs in metastatic neuroblastoma in mice. **a**, Immunoblot analysis of the expression of caspase-8, caspase-9 or caspase-10 in the non-metastatic C1300 neuroblastoma line and two metastatic derivatives, TBJ and NX2, was performed. The caspase-8-expressing L929 cells serve as a reference cell line. **b**, Experimental metastases derived from the NB7C8 neuroblastoma in nude mice were analysed for their expression of mRNA for caspase-2 and caspase-8 by RNase protection analysis. The

parental cell line is shown for comparison. **c**, Experimentally derived metastatic lesions present in the ovary and kidney of two separate mice were assessed for caspase-8 expression protein by immunoblotting (lanes 3 and 5). Cell lines derived from these tumours (lanes 4 and 6) were analysed alongside tumour samples for expression of caspase-8. The input cells (lane 2) and the original caspase-8-negative tumour (lane 1) are shown for reference.

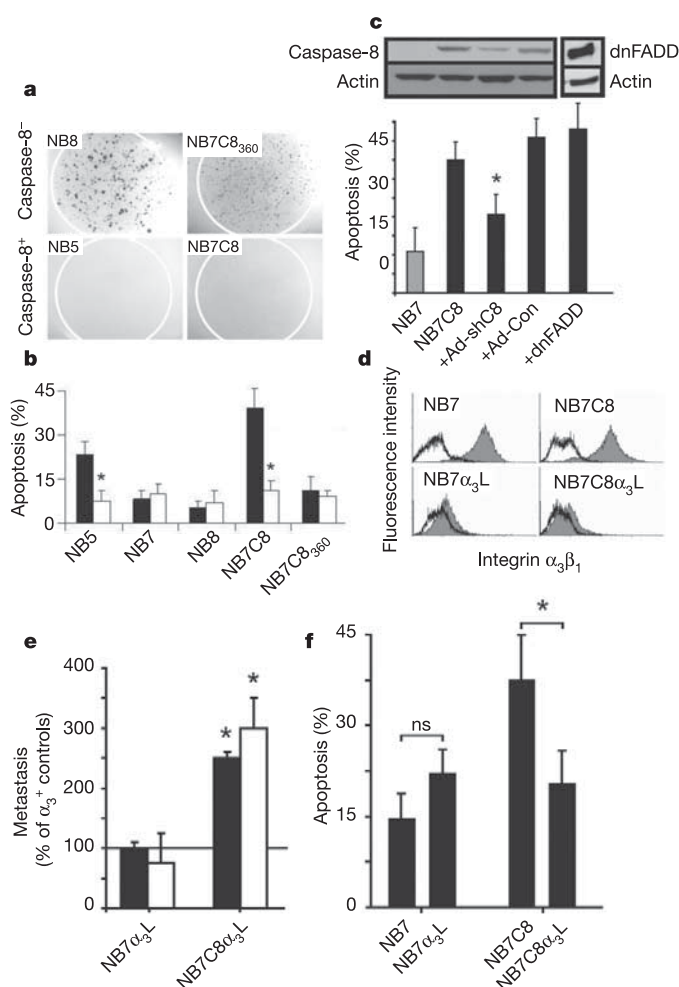


Figure 4 | The induction of apoptosis is dependent on caspase-8 and integrins. **a**, Neuroblastoma cells expressing (NB5, NB7C8) or lacking (NB8, NB7C8360) active caspase-8 were assessed for their capacity to form colonies in collagen gels. **b**, In parallel experiments, neuroblastoma cell apoptosis was assessed at 24 h in the presence (white bars) or absence (black bars) of the caspase inhibitor zVAD (10 μ M). Asterisk, $P < 0.05$. **c**, Neuroblastoma transduced with an adenovirus encoding the shRNA targeting caspase-8 (shC8) or an empty vector virus (control) were evaluated for survival in collagen, as described above. The survival of neuroblastoma cells transduced with a dominant-negative FADD (dnFADD) retrovirus is shown for comparison. The relative expression of caspase-8 or dnFADD was monitored relative to actin in each cell population (see immunoblot inset). Asterisk, $P < 0.05$. **d**, NB7 and NB7C8 neuroblastoma were sorted by fluorescence-activated cell sorting to isolate sublines (NB7- α_3 L and NB7C8- α_3 L) expressing low levels of integrin $\alpha_3\beta_1$. Integrin α_3 expression is shown (filled histogram) relative to controls lacking primary antibody (open histogram) in each case. **e**, The α_3 -deficient cells were assessed for spontaneous metastasis to bone marrow (black bars) or lung (white bars) in the chick embryo, as described in Fig. 2. Results are normalized to metastasis of α_3 -expressing NB7 or NB7C8 cells within the same experiment. Asterisk, $P < 0.05$. **f**, The NB7, NB7- α_3 L, NB7C8 and NB7C8- α_3 L neuroblastoma populations were assessed for survival after incubation in three-dimensional collagen gels for 24 h as described above. Data are means $\pm$ s.e.m. from one of three similar experiments. Asterisk, $P < 0.05$.

METHODS

Cell lines and culture. The neuroblastoma tumour lines were established at St Jude Children's Hospital (NB) or were acquired from ATCC (SKNAS). Caspase-8-negative cells were reconstituted by retroviral transduction with a caspase-8 expression construct, followed by clone-pooling or fluorescence-activated cell sorting for GFP coexpressed from the caspase-8 message (by means of an internal ribosome entry sequence), to select heterogenous polyclonal populations^{3,4}. Caspase-8/dominant-negative FADD double-expressing

cells were generated similarly with yellow fluorescent protein, and sorting was performed to isolate double-positive cell populations (more than 90% of the cells were positive for dominant-negative FADD). Integrin $\alpha_3\beta_1$ 'low' expressing neuroblastoma cells were isolated by multiple rounds of sorting with antibody P1B5 (Chemicon). Collagen survival assays were performed as described previously⁵.

Tumour growth and metastases. For avian tumour studies, 5×10^6 neuroblastoma cells suspended in 40 μ l of complete medium were seeded on 11-day-old chick chorioallantoic membrane^{6,7}. Tumours were left to develop for 7 days and were then resected and weighed. To assess metastasis, lung and bone marrow were harvested from the tumour-bearing embryos and genomic DNA was extracted. Metastasis to lung and bone marrow was determined by polymerase chain reaction (PCR)-based detection of the human Alu sequence using the primers 5'-ACGCCTGTAATCCAGCACTT-3' (sense) and 5'-TCGCCAGGCTG-GAGTGCA-3' (antisense) with chick glyceraldehyde-3-phosphate dehydrogenase-specific primers (sense, 5'-GAGGAAAGGTGCGCTGGTGGATCG-3'; antisense, 5'-GGTGAGGACAAGCAGTGAGGAACG-3') as controls. In both cases, metastasis was assessed by polymerase activation at 95 °C for 2 min followed by 27 cycles at 95 °C for 30 s, 63 °C for 30 s and 72 °C for 30 s. Several experiments were performed in all cases; the s.e.m. is displayed in error bars to indicate variation in the incidence of metastasis observed in the three to eight experiments for each condition. Significance was determined by χ^2 analysis or Fisher's exact test using the total cohort. To generate metastases in SCID CB17 mice, 5×10^6 NB7C8 cells were injected intravenously, twice, on day 0. After 9–12 weeks, disseminated tumours formed principally in the ovary and adrenals, with some hepatic and renal metastases also observed. RNA or protein was isolated from tumours to assess the expression of caspase message by RNase protection analysis or by immunoblotting, as described previously⁸.

Detection of apoptosis *in situ*. Apoptosis was detected in CAM tissue by TUNEL staining of neuroblastoma cells prelabelled with 2 μ M Celltracker red (Molecular Probes). Tumours were fixed *in situ* by the addition of 500 μ l of 1% paraformaldehyde to the surface of the CAM for 15 min while the eggs were incubated on ice. Tumours and surrounding CAM tissue were excised, and tumour/CAM preparations were then washed in PBS containing 1% BSA, permeabilized for 5 min in PBS-T (PBS containing 0.1% Tween 20), then placed in 20 μ g ml⁻¹ Proteinase K in PBS-T for 3 min. Tissue was washed twice in PBS-T and fixed in 1% paraformaldehyde for a further 20 min and finally washed a further three times with PBS-T and blocked for 30 min in PBS containing 1% BSA. Tissues were then TUNEL-stained in accordance with the manufacturer's protocol for tissue (Chemicon) and images were captured with a Bio-Rad 1024 or a Nikon CS-1s confocal microscope. Celltracker-red-labelled tumour cells (1 μ M) were resolved in the red channel, and TUNEL staining in the green channel, by means of independent excitation and image capture. Five images were captured from the periphery of each tumour in a blinded fashion, and the incidence of apoptosis was analysed by emission colocalization with LaserSharp or ImageJ software⁵ and a subsequent *t*-test to assess significance.

Construction and use of caspase-8 short interfering RNA (siRNA) vectors. Two different siRNA sequences targeting caspase-8 were used, one targeting the first codon and the other the beginning of the second death effector domain (Imgenix). The first, new siRNA specific for the 5' end of the caspase-8 open reading frame was synthesized as two complementary DNA oligonucleotides. The sequences were 5'-TTGTGGACTTCAGCAGAAATCTTTGGAAAAA-ATTCGAAAGATTCTGCTGAAGTCCATTTT-3' and 5'-CTAGAAAAA-TGGACTTCAGCAGAAATCTTCGAATTTTCCAAAGATTCTGCTGAA-GTCCA-3'. The underlined sequence corresponds to the targeted sequence in caspase-8 and anneals to the target mRNA beginning with the final nucleotide in the initiating codon. Oligonucleotides were annealed and ligated into mu6pro vector digested with *Bbs*I and *Xba*I. For transfection studies, the mu6pro siRNA (siRNA1) or the Imgenix U6 shRNA to caspase-8 (siRNA2) was co-transfected with a GFP 'marker' plasmid at a 4:1 ratio. GFP-positive cells (about 20% of the total) were scored for apoptosis after a further 24 h. To create the siRNA adenovirus, the siRNA cassette, containing the mouse U6 promoter and the hairpin siRNA sequence, was excised with *Hind*III and *Not*I and ligated into pShuttle (Stratagene). Caspase-8 siRNA adenovirus was produced with the AdEasy Adenoviral Vector System. Cells were infected at a multiplicity of infection of 5 plaque-forming units per cell 2 days before use to establish knockdown of caspase-8.

Received 30 August; accepted 7 October 2005.

1. Brodeur, G. M. Neuroblastoma: biological insights into a clinical enigma. *Nature Rev. Cancer* 3, 203–216 (2003).
2. Takita, J. et al. Allelic imbalance on chromosome 2q and alterations of the caspase 8 gene in neuroblastoma. *Oncogene* 20, 4424–4432 (2001).
3. Teitz, T., Lahti, J. M. & Kidd, V. J. Aggressive childhood neuroblastomas do not

- express caspase-8: an important component of programmed cell death. *J. Mol. Med.* **79**, 428–436 (2001).
4. Teitz, T. *et al.* Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. *Nature Med.* **6**, 529–535 (2000).
 5. Stupack, D. G., Puente, X. S., Boutsaboualoy, S., Storgard, C. M. & Cheresch, D. A. Apoptosis of adherent cells by recruitment of caspase-8 to unligated integrins. *J. Cell Biol.* **155**, 459–470 (2001).
 6. Kim, J., Yu, W., Kovalski, K. & Ossowski, L. Requirement for specific proteases in cancer cell intravasation as revealed by a novel semiquantitative PCR-based assay. *Cell* **94**, 353–362 (1998).
 7. Zijlstra, A. *et al.* A quantitative analysis of rate-limiting steps in the metastatic cascade using human-specific real-time polymerase chain reaction. *Cancer Res.* **62**, 7083–7092 (2002).
 8. Beltinger, C. & Debatin, K. M. Murine models for experimental therapy of pediatric solid tumors with poor prognosis. *Int. J. Cancer* **92**, 313–318 (2001).
 9. DuBois, S. G. *et al.* Metastatic sites in stage IV and IVS neuroblastoma correlate with age, tumour biology, and survival. *J. Pediatr. Hematol. Oncol.* **21**, 181–189 (1999).
 10. Ziegler, M. M., Ishizu, H., Nagabuchi, E., Takada, N. & Arya, G. A comparative review of the immunobiology of murine neuroblastoma and human neuroblastoma. *Cancer* **79**, 1757–1766 (1997).
 11. Lode, H. N. *et al.* Targeted interleukin-2 therapy for spontaneous neuroblastoma metastases to bone marrow. *J. Natl Cancer Inst.* **89**, 1586–1594 (1997).
 12. Barnhart, B. C., Alappat, E. C. & Peter, M. E. The CD95 type I/type II model. *Semin. Immunol.* **15**, 185–193 (2003).
 13. Wang, S. & El-Deiry, W. S. TRAIL and apoptosis induction by TNF-family death receptors. *Oncogene* **22**, 8628–8633 (2003).
 14. Teitz, T. *et al.* Caspase-9 and Apaf-1 are expressed and functionally active in human neuroblastoma tumor cell lines with 1p36 LOH and amplified MYCN. *Oncogene* **21**, 1848–1858 (2002).
 15. Zhao, H., Ross, F. P. & Teitelbaum, S. L. Unoccupied $\alpha_v\beta_3$ integrin regulates osteoclast apoptosis by transmitting a positive death signal. *Mol. Endocrinol.* **19**, 771–780 (2005).
 16. Patan, S., Haenni, B. & Burri, P. H. Implementation of intussusceptive microvascular growth in the chicken chorioallantoic membrane (CAM). *Microvasc. Res.* **53**, 33–52 (1997).
 17. Zijlstra, A. *et al.* Collagenolysis-dependent angiogenesis mediated by matrix metalloproteinase-13 (collagenase-3). *J. Biol. Chem.* **279**, 27633–27645 (2004).
 18. Bredesen, D. E., Mehlen, P. & Rabizadeh, S. Apoptosis and dependence receptors: a molecular basis for cellular addiction. *Physiol. Rev.* **84**, 411–430 (2004).
 19. Bogenrieder, T. & Herlyn, M. Axis of evil: molecular mechanisms of cancer metastasis. *Oncogene* **22**, 6524–6536 (2003).
 20. Bates, R. C., Edwards, N. S. & Yates, J. D. Spheroids and cell survival. *Crit. Rev. Oncol. Hematol.* **36**, 61–74 (2000).
 21. Zahir, N. & Weaver, V. M. Death in the third dimension: apoptosis regulation and tissue architecture. *Curr. Opin. Genet. Dev.* **14**, 71–80 (2004).
 22. Eggert, A. *et al.* Expression of the neurotrophin receptor TrkB is associated with unfavorable outcome in Wilms' tumor. *J. Clin. Oncol.* **19**, 689–696 (2001).
 23. Frisch, S. M. & Ruoslahti, E. Integrins and anoikis. *Curr. Opin. Cell Biol.* **9**, 701–706 (1997).
 24. Jan, Y. *et al.* A mitochondrial protein, Bit1, mediates apoptosis regulated by integrins and Groucho/TLE corepressors. *Cell* **116**, 751–762 (2004).
 25. Stupack, D. G. & Cheresch, D. A. Get a ligand, get a life: integrins, signalling and cell survival. *J. Cell Sci.* **115**, 3729–3738 (2002).
 26. Shivapurkar, N. *et al.* Differential inactivation of caspase-8 in lung cancers. *Cancer Biol. Ther.* **1**, 65–69 (2002).
 27. Pingoud-Meier, C. *et al.* Loss of caspase-8 protein expression correlates with unfavorable survival outcome in childhood medulloblastoma. *Clin. Cancer Res.* **9**, 6401–6409 (2003).
 28. Kim, H. S. *et al.* Inactivating mutations of caspase-8 gene in colorectal carcinomas. *Gastroenterology* **125**, 708–715 (2003).
 29. Tweddle, D. A. *et al.* The p53 pathway and its inactivation in neuroblastoma. *Cancer Lett.* **197**, 93–98 (2003).
 30. Berger, J. C., Vander Griend, D. J., Robinson, V. L., Hickson, J. A. & Rinker-Schaeffer, C. W. Metastasis suppressor genes: From gene identification to protein function and regulation. *Cancer Biol. Ther.* **4**, 805–812 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature. A summary figure is also included.

Acknowledgements In memory of V.J.K. who died 7 May 2004. We thank K. Zhu, J. Creech, J. Grenet, T. Lai and K. Boyd for help. This work was supported by National Cancer Institute grants to D.A.C., D.G.S., P.J.H., V.J.K. and J.M.L., and an NCI CCSG grant and ALSAC support to St Jude Children's Research Hospital. The mU6pro vector was generously provided by D. L. Turner. Caspase-8 shRNA vectors were provided by G. Singh.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.G.S. (dstupack@ucsd.edu) or J.M.L. (jill.lahti@stjude.org).

LETTERS

Mechanochemical analysis of DNA gyrase using rotor bead tracking

Jeff Gore^{1*}, Zev Bryant^{2,3*}†, Michael D. Stone^{2*}†, Marcelo Nöllmann², Nicholas R. Cozzarelli²
& Carlos Bustamante^{1–4}

DNA gyrase is a molecular machine that uses the energy of ATP hydrolysis to introduce essential negative supercoils into DNA^{1–3}. The directionality of supercoiling is ensured by chiral wrapping of the DNA^{4,5} around a specialized domain^{6–9} of the enzyme before strand passage. Here we observe the activity of gyrase in real time by tracking the rotation of a submicrometre bead attached to the side of a stretched DNA molecule¹⁰. In the presence of gyrase and ATP, we observe bursts of rotation corresponding to the processive, stepwise introduction of negative supercoils in strict multiples of two¹¹. Changes in DNA tension have no detectable effect on supercoiling velocity, but the enzyme becomes markedly less processive as tension is increased over a range of only a few tenths of piconewtons. This behaviour is quantitatively explained by a simple mechanochemical model in which processivity depends on a kinetic competition between dissociation and rapid, tension-sensitive DNA wrapping. In a high-resolution variant of our assay, we directly detect rotational pauses corresponding to two kinetic substeps: an ATP-independent step at the end of the reaction cycle, and an ATP-binding step in the middle of the cycle, subsequent to DNA wrapping.

Negative DNA supercoiling is essential *in vivo* to compact the genome, to relieve torsional strain during replication, and to promote local melting for vital processes such as transcript initiation by RNA polymerase^{12,13}. In bacteria, negative supercoiling is achieved through the activity of DNA gyrase, which works against mechanical stresses to drive the genome into an elastically strained configuration. Single-molecule techniques have yielded important insights into the mechanisms of other topoisomerases¹⁴, but have yet to be applied to DNA gyrase.

Gyrase and other type II topoisomerases carry out a complex series of conformational changes resulting in the passage of an intact DNA duplex (the T-segment) through a transient break in another DNA duplex (the G-segment), changing the linking number¹⁵ of the DNA by two¹¹. Gyrase further embellishes this mechanism with a specialized adaptation whereby a chiral DNA wrap is formed before strand passage. The DNA wrap ensures the directionality of topoisomerization and confers on gyrase its unique ability to introduce, rather than merely to relax, DNA supercoils^{4–9}. Wrapping involves a large change in the end-to-end extension of the DNA^{7,16}, and is therefore expected to be sensitive to tension and subject to perturbation in single-molecule assays. The equilibrium properties of DNA wrapped around gyrase or its subdomain have been studied extensively^{4,5,7,8,16,17}, but the dynamics of DNA wrapping remain largely uncharacterized. Other poorly understood aspects of gyrase dynamics include the mechanism of processivity (by which gyrase

can perform many successive strand passages without releasing the DNA substrate), the location of the rate-limiting step for the overall reaction cycle, and the coupling between ATP consumption and supercoil introduction.

To dissect the mechanochemical cycle of DNA gyrase, we have used a method that we developed for measuring torque and changes in twist in a single DNA molecule in real time¹⁰. This rotor bead tracking technique requires a molecular construct containing three distinct chemical modifications (Fig. 1a). Tension is generated in the molecule by pulling at the two ends of the DNA, and the central 'rotor' bead is attached to the middle of the DNA just below an engineered single strand nick, which acts as a free swivel (Fig. 1b). The angle of the rotor bead then reflects changes in twist of the lower DNA segment, and the angular velocity of the bead is proportional to the torque in this segment. Previously we applied tension to the molecule using a laser trap¹⁰, whereas here we use a magnetic tweezers^{18,19} apparatus based on an inverted microscope (Fig. 1b). Tension in the DNA causes changes in linking number to partition into DNA twist^{15,19}, resulting in a torque on the rotor bead. An enzymatic process that changes the linking number by two will cause the rotor bead to spin around twice as the DNA returns to its equilibrium conformation. Thus, the DNA construct serves as a self-regenerating substrate for DNA gyrase.

In the absence of enzyme, the rotor bead fluctuates around a mean angle as a result of thermal noise¹⁰. On addition of *Escherichia coli* gyrase and 1 mM ATP, the rotor bead undergoes bursts of directional rotation (clockwise as viewed from below, as expected for negative supercoiling), separated by periods of inactivity (Fig. 1c). The pauses between bursts become longer as the enzyme concentration is reduced. We infer that a burst of rotation corresponds to the activity of a single enzyme that binds to the DNA, performs one or more catalytic cycles, and then dissociates. Each burst results in an even number of rotations (Fig. 1d), as predicted from the established sign-inversion mechanism of type II topoisomerases¹¹, in which a single catalytic cycle changes the DNA linking number by two.

As tension is varied from 0.35 to 1.3 pN, the activity of gyrase changes markedly (Figs 1c and 2a–c). Unexpectedly, the supercoiling velocity in a processive burst is not a strong function of template tension (Fig. 2a). However, processivity and initiation rate are sensitive to small changes in tension (Figs 1c and 2b, c). At 1 pN, gyrase activity consists almost exclusively of isolated single enzymatic cycles, whereas at lower force the bursts of rotation increase in length (Figs 1c and 2b). Burst initiation is also strongly suppressed by force: the waiting times between bursts increase by more than two orders of magnitude over a 0.5-pN range of tensions (Figs 1c and 2c).

¹Department of Physics, University of California, Berkeley, California 94720, USA. ²Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. ³Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. ⁴Howard Hughes Medical Institute, USA. †Present addresses: Department of Biochemistry, Stanford University, Stanford, California 94305, USA (Z.B.); Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138, USA (M.D.S.).

*These authors contributed equally to this work.

A simple mechanochemical model quantitatively explains these experimental results (Fig. 3). We consider a state ('gyrase-DNA') in which DNA is bound to the enzyme but not yet fully wrapped. We expect this state to be vulnerable to rapid dissociation because of its limited protein-DNA-binding interface and by analogy to yeast topoisomerase II, which dissociates from a 40-bp DNA segment at

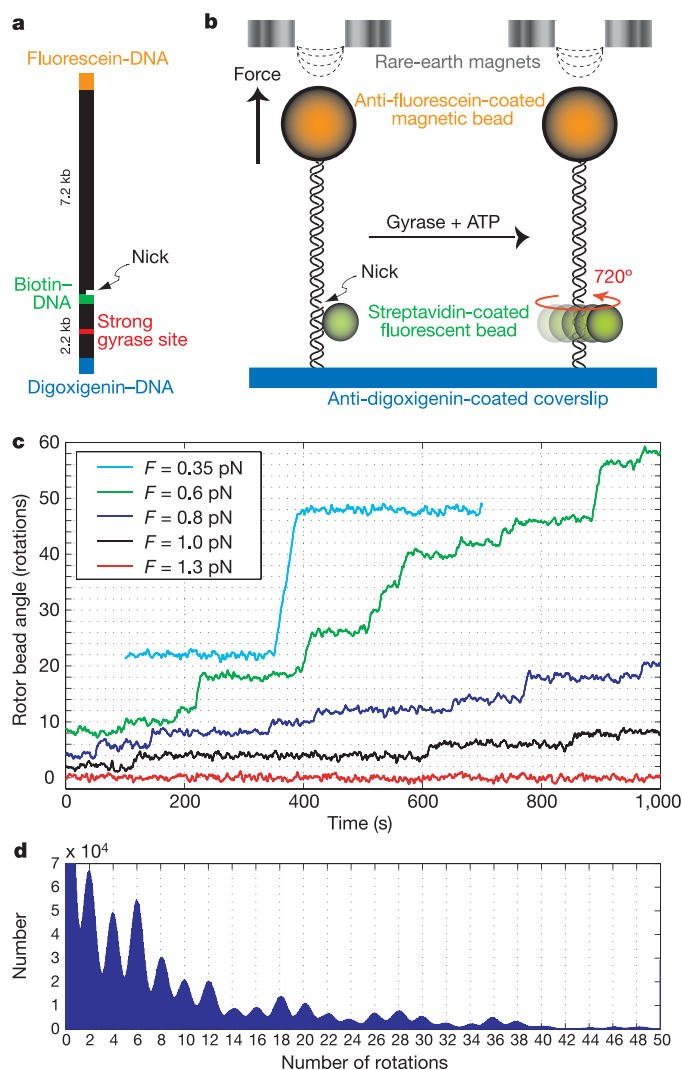


Figure 1 | Experimental design and single-molecule observations of gyrase activity. **a**, The molecular construct contains three distinct attachment sites and a site-specific nick, which acts as a swivel. A strong gyrase site was engineered into the lower DNA segment²⁹. **b**, Molecule and bead assemblies were constructed in parallel in a flow chamber and assayed by using an inverted microscope equipped with permanent magnets. Each molecule was stretched between the glass coverslip and a 1- μ m magnetic bead, and a 530-nm diameter fluorescent rotor bead was attached to the central biotinylated patch. In the presence of gyrase and ATP, the rotor bead underwent bursts of rotation due to the enzymatic activity of individual gyrase enzymes acting on the DNA segment below the rotor bead. **c**, Plot of the rotor bead angle as a function of time (averaged over a 2-s window), showing bursts of activity due to diffusional encounters of individual gyrase enzymes. The activity of the enzyme is strongly dependent on tension. With the exception of the one at 0.35 pN, the traces shown were taken in the same chamber with a single concentration of gyrase; thus, the differences in burst density reflect force-dependent initiation rates. **d**, Histogram of the pairwise difference distribution function summed over 11 traces of 15–20 min (averaged over a 4-s window) at forces of 0.6–0.8 pN. The spacing of the peaks indicates that each catalytic cycle of the enzyme corresponds to two full rotations of the rotor bead, as expected for a type II topoisomerase such as DNA gyrase.

~100 Hz (F. Mueller-Planitz and D. Herschlag, personal communication). According to our model, the gyrase-DNA complex undergoes a kinetic competition between DNA wrapping (which in the presence of ATP commits the enzyme to a productive cycle) and dissociation (which terminates a processive burst). At the end of each cycle, gyrase returns to the vulnerable state and must repeat its kinetic choice. In the absence of mechanical stresses, wrapping is very rapid, tending to outcompete dissociation and leading to processive supercoiling. Increased tensions slow down the DNA wrapping step, lengthening the time spent in the vulnerable state and increasing the probability of dissociation.

The average number of enzymatic cycles $\langle n \rangle$ in each burst of activity is given by $\langle n \rangle = 1/(1 - P_{\text{cycle}})$, where

$$P_{\text{cycle}}(F) = \frac{k_{\text{wrap}}(F)}{k_{\text{wrap}}(F) + k_{\text{off}}}$$

is the probability of the gyrase-DNA complex committing to a productive catalytic cycle. Assuming a simple transition state model²⁰ with a distance to the transition state of Δx_t , the rate of wrapping is given by $k_{\text{wrap}} = k_{\text{wrap},0} \exp(-F\Delta x_t/k_B T)$, where $k_B T = 4.1$ pN nm is the thermal energy of the bath and F is the

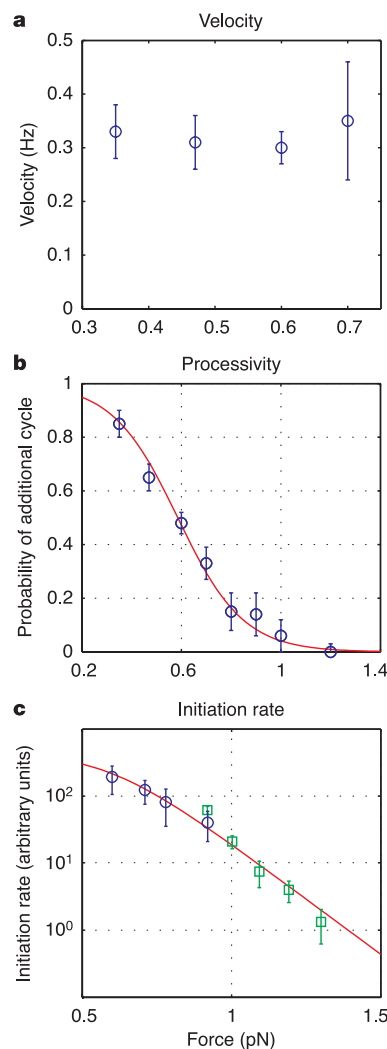


Figure 2 | Modulation of gyrase activity by DNA tension. The velocity (**a**) within a burst is insensitive to DNA tension, but both the processivity (**b**) and initiation rate (**c**) decrease rapidly as DNA tension increases. In **c**, tension-dependent initiation rates were measured in two independent experiments after the introduction of 10 nM (green squares) or 5 nM (blue circles) gyrase. Error bars indicate the s.e.m.

force applied to the magnetic bead. In our model, the observed burst frequency $k_{\text{init}} = [\text{gyrase}]k_{\text{on}}P_{\text{cycle}}$ decreases at higher force because of an increased number of 'invisible' gyrase–DNA binding events in which the enzyme dissociates before carrying out a single cycle. In Fig. 2b, c we fit $P_{\text{cycle}}(F)$ and $k_{\text{init}}(F)$ using the two parameters $\Delta x_t = 31 \pm 3$ nm and $k_{\text{wrap},0}/k_{\text{off}} = 85 \pm 30$ (red line; see also Methods and Supplementary Note). This single pair of parameter values yields good fits to both processivity and initiation rate data, supporting the idea that a kinetic competition between wrapping and dissociation underlies both motor properties. The measured transition state distance of ~ 31 nm is close to the total expected decrease in DNA extension on wrapping of ~ 120 bp (~ 40 nm)¹⁶.

Unlike DNA wrapping, the burst velocity is insensitive to tension (Figs 1c and 2a) and must be limited by the rate of some other, force-insensitive step. Having used mechanical perturbation²¹ to characterize the wrapping step, we employed a complementary strategy to characterize the rate-determining step: direct observation of pauses in the reaction cycle^{22,23}. We improved the time resolution of our assay by reducing the size of the rotor bead and shortening the torque-bearing DNA segment (see Methods). High-resolution traces (Fig. 4a) contain many distinct pauses within processive bursts. We tried to locate the position along the repeating 720° reaction coordinate at which the main pause in the cycle occurs. Most of the pronounced pauses occurred after two full rotations of the rotor bead (Fig. 4a and Supplementary Fig. 1). The remaining pauses occurred after about one rotation of the rotor bead (Fig. 4a, asterisks). A histogram of pause durations (Supplementary Fig. 1) confirms that the main pause (~ 1.5 s) is at the two-rotation mark, but suggests the possibility of a shorter (~ 0.8 s) pause in the middle of the cycle.

The positions of intraburst pauses indicate that the principal rate-determining step occurs near the beginning or the end of the reaction cycle. DNA wrapping lies at the beginning of the cycle, but we have already shown that this step is not rate-limiting. We therefore conclude that the rate-limiting step occurs at the end of the cycle (denoted k_{rl} in Fig. 3). The shorter, mid-cycle pause might correspond to an intermediate following DNA wrapping and preceding strand passage. Exit from this intermediate is expected to depend on ATP, because ATP binding is required for strand passage^{24,25} but not for DNA wrapping⁵ (Fig. 3). To directly investigate the role of ATP in the cycle, we carried out high-resolution assays at varying ATP concentrations. In the complete absence of ATP, the cycle should

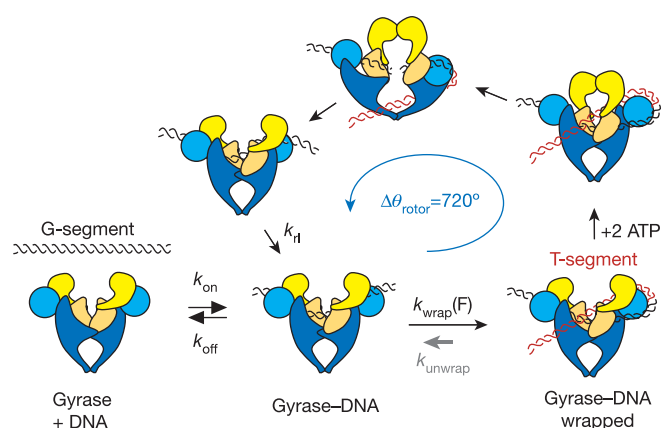


Figure 3 | Proposed mechanochemical model. From the state labelled 'gyrase–DNA', there is a kinetic competition between DNA wrapping and dissociation. Wrapping is strongly inhibited by DNA tension. After wrapping and ATP binding, the enzyme is committed to a full catalytic cycle in which two negative supercoils are introduced to the DNA, causing the rotor bead to spin by $\Delta\theta = 720^\circ$. At saturating concentrations of ATP, unwrapping (small arrow labelled k_{unwrap}) is negligible; however, see Fig. 4b, c.

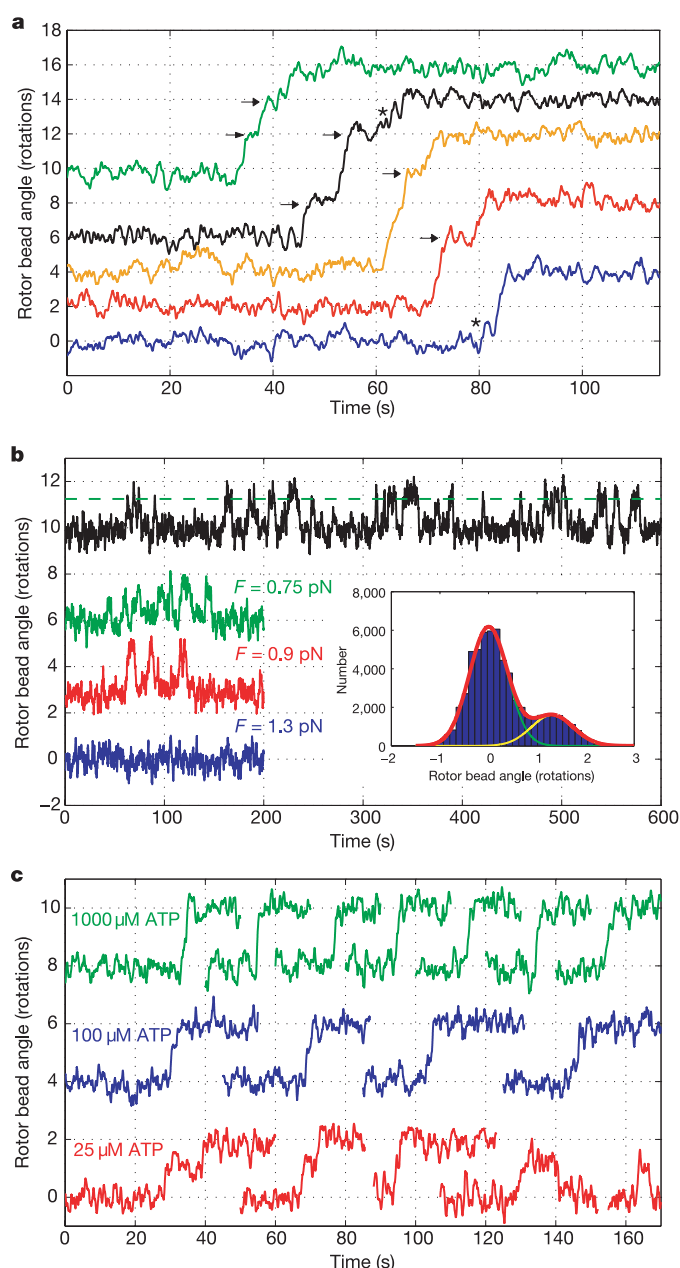


Figure 4 | Gyrase activity observed at high resolution. **a**, High-resolution traces ($F \approx 0.8$ pN) at 1 mM ATP show that the dominant pause in the catalytic cycle occurs at the two-rotation mark, corresponding to either the beginning or the end of the cycle (arrows). Pauses were less frequently observed in the middle of the catalytic cycle (asterisks). Traces were averaged over a 300-ms window. Burst velocity in the high-resolution assay was 0.38 ± 0.04 Hz (mean $\pm$ s.e.m.), which was not significantly faster than in the low-resolution assay (0.32 ± 0.03 Hz, Fig. 2a). **b**, In the absence of ATP, the rotor bead angle alternates between two values, as expected for reversible DNA wrapping (black trace, $F = 0.9$ pN). The wrapped state corresponds to a change in the angle of the rotor bead of ~ 1.3 rotations (broken green line), as shown by a double-gaussian fit to the histogram of rotor bead angles for this trace (inset). Increasing the DNA tension from 0.75 pN (green trace) to 1.3 pN (blue trace) strongly inhibits wrapping. **c**, Fine structure of isolated enzymatic cycles at several ATP concentrations. The mid-cycle pause at about the one-rotation mark becomes more pronounced as the ATP concentration is lowered, revealing an ATP-binding step subsequent to DNA wrapping (green traces, 1 mM ATP; blue traces, 100 μ M ATP; red traces, 25 μ M ATP). At 25 μ M ATP, unwrapping often occurs before the cycle can be completed (final red traces).

halt at the wrapped intermediate ('gyrase-DNA wrapped' in Fig. 3), after roughly one rotation of the rotor bead⁵. The wrapped state should persist until a rare unwrapping event occurs (k_{unwrap} in Fig. 3), returning the rotor bead to its original position.

As predicted from this model, we observed reversible rotations of the rotor bead when low concentrations of gyrase were introduced in the absence of ATP (Fig. 4b). The change in angle of the rotor bead on wrapping is ~ 1.3 rotations (Fig. 4b, inset), slightly larger than the value (0.8 rotations) predicted from bulk topoisomerase I protection assays¹⁷. This discrepancy may be due to the difficulty of estimating the number of bound enzymes in bulk experiments. The wrapped state is long-lived (with a mean lifetime of 7 s at $F = 0.9$ pN, $N = 66$), in accordance with a primary assumption of our mechanochemical model: namely, the wrapped and unwrapped states are not in rapid equilibrium. As expected from our model, wrapping events are strongly suppressed by tension (Fig. 4b).

At subsaturating ATP concentrations, the cycle can progress to completion only on ATP binding. Lowering the ATP concentration should therefore lengthen the time spent in a wrapped intermediate state. This prediction is confirmed by observations of single enzymatic cycles at varying ATP concentrations in the high-resolution assay (Fig. 4c). At saturating ATP, a pause at the one-rotation mark is only rarely observed. As the ATP concentration is reduced, the mid-cycle pause becomes clearly visible. At 25 μM ATP, unwrapping often occurs before the catalytic cycle can be completed, reflecting a kinetic competition between unwrapping and ATP binding (Fig. 4c, final red traces).

Chiral DNA wrapping enables gyrase to introduce essential negative supercoils into the bacterial genome¹⁸. We have found that this wrapping step is rapid and exquisitely sensitive to tension in the DNA. Forces of ~ 1 pN (small in comparison to the >20 pN stall forces of DNA-based motors such as RNAP²⁰ and FtsK²⁶) are sufficient to inhibit DNA wrapping by a factor of $\sim 1,000$. Rapid wrapping is essential for processivity, illustrating a general design principle for processive motors^{27,28}: any on-pathway state that is vulnerable to dissociation requires a mechanism for rapid exit from this state into a more tightly bound configuration.

Because DNA wrapping is chiral, it is expected to be sensitive to torque as well as tension. Evidence for torque sensitivity can be seen in the difference in processivity between the high- and low-resolution assays: high-resolution traces taken over a range of forces (0.65–0.9 pN) show enhanced processivity (with the tension at the midpoint $P_{\text{cycle}}(F) = 0.5$ shifted upwards by ~ 0.15 pN), suggesting that wrapping is decelerated by the small amount of torque (~ 3 pN nm) that accumulates owing to rotational drag in the low-resolution assay.

Direct detection of pauses in the supercoiling reaction by the high-resolution rotor bead tracking experiments presented here has shown that the rate-limiting step lies at the end of the cycle. Identifying the chemical and conformational nature of this 'reset' step presents a future challenge for biophysical investigations of DNA gyrase. A second, mid-cycle pause is directly observed at low ATP concentrations and corresponds to an intermediate awaiting ATP binding subsequent to DNA wrapping. Further studies of the drug and nucleotide dependence of the primary and secondary pauses²² will illuminate the important issue of coupling between ATP turnover and DNA supercoiling.

METHODS

Molecular constructs and beads. Modified DNA constructs for rotor bead tracking were prepared as described¹⁰, with the torsionally constrained DNA segment replaced by a 2.2 kb (low-resolution) or 1.1 kb (high-resolution) fragment derived from pMP1000 (a gift from P. Higgins), a plasmid containing the nuB 74 variant of the strong gyrase site from μ phage DNA²⁹. We coated 1- μm magnetic beads (Dyna) with rabbit anti-fluorescein (Molecular Probes). We used streptavidin-coated 'Dragon Green' 0.53- μm rotor beads (Bangs Labs) or avidin-coated 'Yellow' nominal 0.46- μm rotor beads (SpheroTech) fluorescent particles.

Chamber preparation. Flow chambers were prepared as described¹⁰ and incubated for ~ 4 h with 0.2 mg ml⁻¹ of anti-digoxigenin (Roche) in PBS followed by >8 h with a passivation buffer made by mixing equal volumes of 10 mg ml⁻¹ of bovine serum albumin (BSA; NEB) and 80 mM Tris-HCl (pH 8.0), 1 M NaCl, 0.04% azide and 0.4% Tween-20. Molecule-bead assemblies were generated by successively incubating the chamber with DNA (~ 4 pM, 30 min), fluorescent beads (2 h) and magnetic beads (1 h) in binding buffer (40 mM Tris-HCl, 500 mM NaCl, 0.02% azide, 0.2% Tween-20 and 500 μg ml⁻¹ of BSA; pH 8.0).

Microscopy. Single-molecule experiments were conducted on a modified Axiovert 100A inverted microscope (Zeiss) with a pair of neodymium-iron-boron magnets (Radio Shack) suspended above the flow chamber. The force on the magnetic bead was varied by raising or lowering the magnets. Brightfield images of magnetic beads or fluorescence images of rotor beads were imaged on an IXON electron-multiplying CCD camera (Andor). Forces ($\pm 20\%$) were determined by measuring the transverse fluctuations of the magnetic bead¹⁹. Movies of rotor beads were recorded at 40 Hz (100 Hz for the high-resolution experiments) for subsequent video analysis. The centroid of the fluorescent bead was determined with an accuracy of ~ 10 nm per frame, which translates into an error in the angle of $\sim 3^\circ$ (0.05 rad). In the low-resolution assay, the variance in the bead angle was 8 rad^2 with a relaxation time of 1.5 s. In the high-resolution assay, the variance was reduced to 4.5 rad^2 with a relaxation time of ~ 400 ms. Distinguishing between two angles of the rotor bead that are one rotation apart therefore requires ~ 1 s.

Experimental protocol. After flushing the channel with binding buffer, magnets were placed above the chamber and ~ 5 nM gyrase was added to the chamber in reaction buffer (35 mM Tris-HCl, 24 mM potassium glutamate, 4 mM MgCl₂, 2 mM dithiothreitol, 100 μg ml⁻¹ of BSA and 0.2 mM spermidine; pH 7.6) containing 1 mM ATP. All saturating ATP data were taken with enzyme provided by A. Maxwell and all other data were taken with enzyme provided by J. Berger. All experiments were done at room temperature ($23 \pm 2^\circ\text{C}$). The ATP concentration in the low ATP experiments was kept constant by adding an ATP regeneration system containing 10 mM phosphocreatine and 1.23 μM creatine phosphokinase. Activity in the chamber (burst density) for a given concentration of gyrase was variable and fell off over the course of hours, probably owing to inactivation and sticking of the enzyme to the chamber walls. To control for effective gyrase concentration, comparisons of burst (Fig. 2c) and wrapping (Fig. 4b) initiation rates were done in a single chamber by alternating forces without reintroducing enzyme.

Data analysis. Burst velocities were measured by a piecewise linear fit to the raw angle data for each burst containing four or more catalytic cycles. Mean velocities were calculated by summing the total number of cycles in all the bursts (typically $N \approx 50$ at each force) and dividing by the total time of all the bursts. Processive bursts in the low-resolution assay could be analysed only for forces between 0.3 pN and 0.7 pN. Forces below 0.3 pN are inaccessible because lateral fluctuations of the DNA interfere with tracking of the rotor bead. At forces above 0.7 pN, processive bursts become too rare to measure a velocity.

The average burst length ($\langle n \rangle$) was calculated by adding the total number of enzymatic cycles completed at a given force and dividing by the total number of bursts at that force. Processivity was then plotted as $P_{\text{cycle}} = 1 - 1/\langle n \rangle$. Initiation rate as a function of force was measured in two separate experiments (using 5 nM or 10 nM gyrase) for the low-force and high-force regimes. Initiation rates for each data set were multiplied by a fit scaling factor (related to the effective concentration) before plotting.

Received 13 June; accepted 11 October 2005.

- Gellert, M., Mizuuchi, K., O'Dea, M. H. & Nash, H. A. DNA gyrase: an enzyme that introduces superhelical turns into DNA. *Proc. Natl Acad. Sci. USA* **73**, 3872–3876 (1976).
- Champoux, J. J. DNA topoisomerases: structure, function, and mechanism. *Annu. Rev. Biochem.* **70**, 369–413 (2001).
- Corbett, K. D. & Berger, J. M. Structure, molecular mechanisms, and evolutionary relationships in DNA topoisomerases. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 95–118 (2004).
- Liu, L. F. & Wang, J. C. DNA-DNA gyrase complex: the wrapping of the DNA duplex outside the enzyme. *Cell* **15**, 979–984 (1978).
- Liu, L. F. & Wang, J. C. *Micrococcus luteus* DNA gyrase: active components and a model for its supercoiling of DNA. *Proc. Natl Acad. Sci. USA* **75**, 2098–2102 (1978).
- Reece, R. J. & Maxwell, A. The C-terminal domain of the *Escherichia coli* DNA gyrase A subunit is a DNA-binding protein. *Nucleic Acids Res.* **19**, 1399–1405 (1991).
- Corbett, K. D., Shultzaberger, R. K. & Berger, J. M. The C-terminal domain of DNA gyrase A adopts a DNA-bending β -pinwheel fold. *Proc. Natl Acad. Sci. USA* **101**, 7293–7298 (2004).

8. Ruthenburg, A. J., Graybosch, D. M., Huetsch, J. C. & Verdine, G. L. A superhelical spiral in the *Escherichia coli* DNA gyrase A C-terminal domain imparts unidirectional supercoiling bias. *J. Biol. Chem.* **280**, 26177–26184 (2005).
9. Kampranis, S. C. & Maxwell, A. Conversion of DNA gyrase into a conventional type II topoisomerase. *Proc. Natl Acad. Sci. USA* **93**, 14416–14421 (1996).
10. Bryant, Z. *et al.* Structural transitions and elasticity from torque measurements on DNA. *Nature* **424**, 338–341 (2003).
11. Brown, P. O. & Cozzarelli, N. R. A sign inversion mechanism for enzymatic supercoiling of DNA. *Science* **206**, 1081–1083 (1979).
12. Levine, C., Hiasa, H. & Mariani, K. J. DNA gyrase and topoisomerase IV: biochemical activities, physiological roles during chromosome replication, and drug sensitivities. *Biochim. Biophys. Acta* **1400**, 29–43 (1998).
13. Wang, J. C. Cellular roles of DNA topoisomerases: a molecular perspective. *Nature Rev. Mol. Cell. Biol.* **3**, 430–440 (2002).
14. Charvin, G., Strick, T. R., Bensimon, D. & Croquette, V. Tracking topoisomerase activity at the single-molecule level. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 201–219 (2005).
15. Bates, A. D. & Maxwell, A. *DNA Topology* (Oxford Univ. Press, Oxford, 2005).
16. Heddle, J. G., Mittelheiser, S., Maxwell, A. & Thomson, N. H. Nucleotide binding to DNA gyrase causes loss of DNA wrap. *J. Mol. Biol.* **337**, 597–610 (2004).
17. Kampranis, S. C., Bates, A. D. & Maxwell, A. A model for the mechanism of strand passage by DNA gyrase. *Proc. Natl Acad. Sci. USA* **96**, 8414–8419 (1999).
18. Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
19. Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. & Croquette, V. The elasticity of a single supercoiled DNA molecule. *Science* **271**, 1835–1837 (1996).
20. Wang, M. D. *et al.* Force and velocity measured for single molecules of RNA polymerase. *Science* **282**, 902–907 (1998).
21. Keller, D. & Bustamante, C. The mechanochemistry of molecular motors. *Biophys. J.* **78**, 541–556 (2000).
22. Yasuda, R., Noji, H., Yoshida, M., Kinosita, K. Jr & Itoh, H. Resolution of distinct rotational substeps by submillisecond kinetic analysis of F₁-ATPase. *Nature* **410**, 898–904 (2001).
23. Uemura, S., Higuchi, H., Olivares, A. O., De La Cruz, E. M. & Ishiwata, S. Mechanochemical coupling of two substeps in a single myosin V motor. *Nature Struct. Mol. Biol.* **11**, 877–883 (2004).
24. Roca, J. & Wang, J. C. The capture of a DNA double helix by an ATP-dependent protein clamp: a key step in DNA transport by type II DNA topoisomerases. *Cell* **71**, 833–840 (1992).
25. Baird, C. L., Harkins, T. T., Morris, S. K. & Lindsley, J. E. Topoisomerase II drives DNA transport by hydrolyzing one ATP. *Proc. Natl Acad. Sci. USA* **96**, 13685–13690 (1999).
26. Pease, P. J. *et al.* Sequence-directed DNA translocation by purified FtsK. *Science* **307**, 586–590 (2005).
27. Block, S. M. Leading the procession: new insights into kinesin motors. *J. Cell Biol.* **140**, 1281–1284 (1998).
28. Vale, R. D. Myosin V motor proteins: marching stepwise towards a mechanism. *J. Cell Biol.* **163**, 445–450 (2003).
29. Pato, M. L., Howe, M. M. & Higgins, N. P. A DNA gyrase-binding site at the center of the bacteriophage μ genome is required for efficient replicative transposition. *Proc. Natl Acad. Sci. USA* **87**, 8716–8720 (1990).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Crisone, P. Arimondo, A. Vologodskii, A. Edelstein, S. Mittelheiser, A. Schoeffler, and F. Mueller-Planitz for discussions; A. Maxwell and J. Berger for enzymes; P. Higgins for plasmids; and C. Hodges, M. Le and D. Jennings for technical assistance. J.G. acknowledges funding from the Hertz Foundation. This work was supported by the NIH and the DOE.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.B. (carlos@alice.berkeley.edu) or N.R.C. (ncozzare@berkeley.edu).

RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP

Sophie Dumont^{1*}, Wei Cheng^{2*}, Victor Serebrov⁵, Rudolf K. Beran⁵, Ignacio Tinoco Jr³, Anna Marie Pyle⁵ & Carlos Bustamante^{1,2,3,4,6}

Helicases are a ubiquitous class of enzymes involved in nearly all aspects of DNA and RNA metabolism. Despite recent progress in understanding their mechanism of action, limited resolution has left inaccessible the detailed mechanisms by which these enzymes couple the rearrangement of nucleic acid structures to the binding and hydrolysis of ATP^{1,2}. Observing individual mechanistic cycles of these motor proteins is central to understanding their cellular functions. Here we follow in real time, at a resolution of two base pairs and 20 ms, the RNA translocation and unwinding cycles of a hepatitis C virus helicase (NS3) monomer. NS3 is a representative superfamily-2 helicase essential for viral replication³, and therefore a potentially important drug target⁴. We show that the cyclic movement of NS3 is coordinated by ATP in discrete steps of 11 ± 3 base pairs, and that actual unwinding occurs in rapid smaller substeps of 3.6 ± 1.3 base pairs, also triggered by ATP binding, indicating that NS3 might move like an inchworm^{5,6}. This ATP-coupling mechanism is likely to be applicable to other non-hexameric helicases involved in many essential cellular functions. The assay developed here should be useful in investigating a broad range of nucleic acid translocation motors.

NS3 is a key component of the hepatitis C virus (HCV) RNA replication machinery and lies in a membrane-bound complex with other proteins^{7,8}. NS3 is an NTPase with 3' to 5' helicase activity^{9,10}, and it has been structurally characterized in various contexts¹¹. We have developed a single-molecule^{12–16} assay for directly following the movement of full-length NS3 on its RNA substrate. Specifically, we use optical tweezers to apply a constant tension between two beads attached to the ends of a 60-base-pair (bp) RNA hairpin (Fig. 1a) and monitor the end-to-end distance change of the RNA as it is unwound by NS3. To establish the basis for interpretation of the enzymatic activity, we initially characterize the mechanical unfolding of the substrate in the absence of enzyme. The substrate unfolds at a force of 20.4 ± 0.2 pN (Fig. 1b). When the substrate is held at a constant force below 19 pN with the instrument's force feedback mechanism, no unfolding takes place over periods of several minutes. Substrate unfolding at external forces below 19 pN must therefore be helicase-catalysed.

To follow NS3-catalysed unwinding, we flow NS3 (1–90 nM) and ATP (0.05–1 mM) together in buffer U (see Methods). We then hold the RNA substrate at a constant force of between 5 and 17 pN. NS3 loads on its substrate by means of a 3' single-stranded RNA loading site. As NS3 unwinds the hairpin, the bead separation increases so as to hold the force on the molecule constant (Fig. 1b). The bead separation can be converted, at that force, into the number of RNA base pairs unwound as a function of time by using the worm-like-chain model of RNA elasticity¹⁷. The molecular

geometry results in the release of two nucleotides (nt) for each base pair unwound, thereby amplifying the unwinding signal. The hairpin loop facilitates substrate reformation, allowing several unwinding traces to be collected with each RNA substrate. Unless otherwise noted, data are collected at $22 \pm 1^\circ\text{C}$, 20 nM NS3, 1 mM ATP and 17 pN.

A rich variety of strictly ATP-dependent NS3 behaviours is observed throughout the unwinding traces obtained ($N > 1000$; Supplementary Fig. 1). The extension increases in sharp bursts of rapid strand separation (steps) followed by periods of constant extension (pauses) (Fig. 1c). A histogram of pairwise distances among all extensions of a given trace reveals that the distances between pauses are not randomly distributed but occur with well-defined periodicity (Fig. 1d). A Fourier analysis over each histogram yields an apparent unwinding step size of 11 ± 3 bp (Supplementary Discussion) over more than 100 traces under identical conditions.

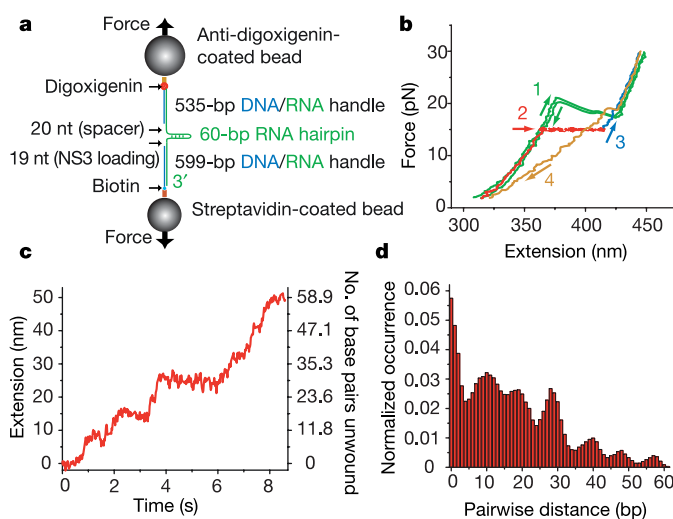


Figure 1 | Assay with optical tweezers for assessing the mechanistic cycle of NS3. **a**, Experimental design and attachment of the RNA substrate. Not to scale. **b**, Stages of an unwinding experiment: the substrate is first unfolded and refolded with mechanical force (green), next brought to a constant force chosen between 5 and 17 pN to monitor NS3-catalysed unwinding (red), then brought to 30 pN to probe its state (blue), and finally brought to 2 pN to allow refolding (yellow; 50% of traces, as this one, display incomplete substrate refolding because of NS3 binding). **c**, Representative trace of extension against time unwinding (15 pN, from **b**). **d**, Pairwise distance distribution for the unwinding trace in **c** (1-bp bins).

¹Biophysics Graduate Group, ²Molecular and Cell Biology Department, ³Chemistry Department, ⁴Physics Department and Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. ⁵Department of Molecular Biophysics and Biochemistry and Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520, USA. ⁶Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

*These authors contributed equally to this work.

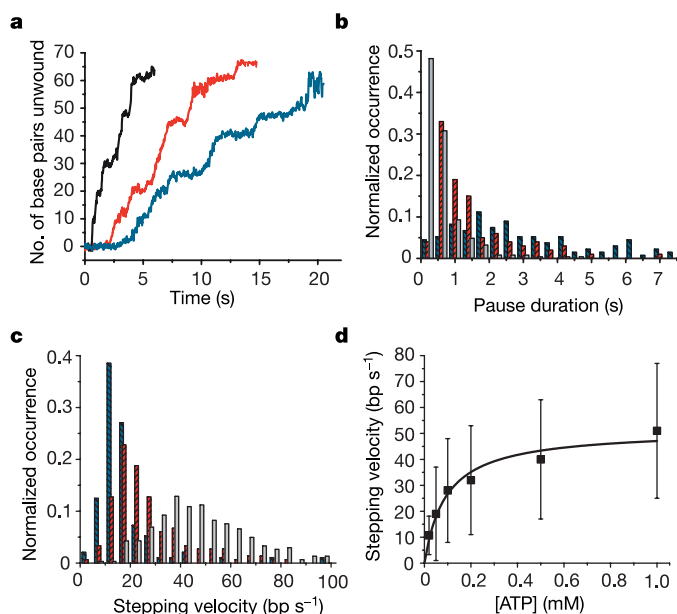


Figure 2 | [ATP] affects both NS3 pauses and steps. **a**, Representative traces of extension against time unwinding with 5 nM NS3 at 1 mM ATP (black), 0.1 mM ATP (red) and 0.05 mM ATP (blue). Traces are displaced along the time axis to avoid overlap. **b**, Histograms of pause durations (0.4-s bins) at 1 mM ATP (grey; 102 traces), 0.1 mM ATP (red; 52 traces) and 0.05 mM ATP (blue; 50 traces). **c**, Histograms of stepping velocities (5 bp s⁻¹ bins) for the data used in **b**. **d**, Stepping velocity as a function of [ATP] with Michaelis-Menten fit. Error bars show s.d.

This step size appears to be intrinsic to NS3 because it is independent of substrate mechanical unfolding pattern and sequence, applied force, ATP concentration and NS3 concentration (Supplementary Figs 2 and 3)¹⁸. We observe rarer apparent backward^{15,16} steps by NS3, corresponding to stepwise refolding of the substrate after complete or partial unwinding (Supplementary Fig. 1c, d). It is possible that NS3 moves backwards (5' to 3') on the same strand¹⁹, or continues forwards (3' to 5') on the other strand, for example after turning around the tetraloop (Supplementary Fig. 1c). The backward step size, 12 ± 3 bp, is statistically indistinguishable from that of forward unwinding steps.

The direct observation of helicase unwinding steps confirms the cyclical movement deduced from bulk measurements of several helicases (for example refs 1, 20, 21). More importantly, it gives direct access to individual mechanistic cycles of a helicase during translocation and unwinding and their coupling to ATP binding. To investigate the role of ATP in coordinating the mechanistic cycle of NS3, we measured the dependence of the cycle on [ATP]. We varied [ATP] below and above the Michaelis-Menten constant, K_m ($160 \pm 6 \mu\text{M}$ for the helicase domain of NS3 (ref. 22)). We observe two effects of [ATP] on the stepping behaviour of NS3 (Fig. 2a). First, the mean pause duration of NS3 decreases with increasing [ATP], from 3.9 s at 0.05 mM ATP to 0.6 s at 1 mM ATP (Fig. 2b), indicating that exit from a pause requires ATP binding. The shapes of the pause duration distributions obtained between 0.05 and 1 mM ATP (Supplementary Fig. 4) indicate that ATP binding might not be the sole rate-limiting event required for pause exit. Rather, the data indicate a possible two-step kinetic mechanism to exit from the pause state, only one of which involves ATP binding. Globally fitting the data to such a mechanism reveals the rates of ATP binding ($k_b = (9.9 \pm 1.0) \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$)²³ and of an [ATP]-independent step² ($k_o = 1.9 \pm 0.1 \text{ s}^{-1}$) during pause. Second, the stepping velocity of NS3 between two pauses (that is, the slope of the steps in Fig. 2a) increases with [ATP], from $19 \pm 18 \text{ bp s}^{-1}$ at 0.05 mM ATP to $51 \pm 26 \text{ bp s}^{-1}$ at 1 mM ATP (Fig. 2c). Dependence of the stepping

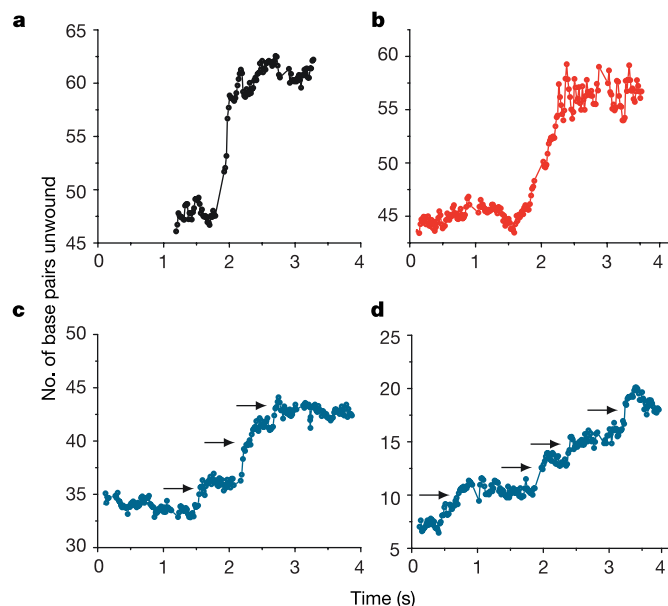


Figure 3 | NS3 steps are composed of substeps. Number of base pairs unwound over time for representative steps observed with 5 nM NS3 at 1 mM ATP (**a**), 0.1 mM ATP (**b**) and 0.05 mM ATP (**c** and **d**, displaying different numbers of substeps). Arrows point to substeps. About 90% of the steps observed at 1 mM ATP show no visible substeps within the step.

velocity on [ATP] implies that each step observed here is not one elementary event with a constant velocity but is made up of substeps, each of which requires ATP binding. The dependence of the stepping velocity of NS3 on [ATP] (Fig. 2d) follows Michaelis-Menten kinetics with $V_{\text{max}} = 51 \pm 3 \text{ bp s}^{-1}$ and $K_m = 93 \pm 21 \mu\text{M}$. Thus, ATP binding must also take place during each step, strongly indicating that each unwinding step may be composed of substeps, which is consistent with the free energy of hydrolysis of one ATP molecule not being sufficient to melt 11 bp.

Indeed, more than 50% of unwinding traces collected at 0.05 mM ATP show substeps that appear within the 11 ± 3 -bp step (Fig. 3, arrows). Substeps are visible even under saturating [ATP] conditions, although with lower frequency. The fact that the motor is observed to take substeps in a wide range of [ATP] indicates that its fundamental mechanism of operation is conserved through all [ATP]. Increasing [ATP] simply decreases the time spent between substeps and thus the probability of observing them. Most substeps are between 2 and 5 bp (3.6 ± 1.3 bp; Supplementary Fig. 5), indicating the possible existence of three substeps per step on average. A Poisson analysis of stepping times at 0.05 and 0.1 mM ATP confirms that each step is indeed composed of three substeps (Supplementary Fig. 5c). The presence of substeps suggests that in a full mechanistic cycle, unwinding is achieved not in a single motion by the helicase but in discrete smaller motions coordinated and triggered by ATP binding.

How many translocation and unwinding cycles can NS3 perform before dissociating from its substrate? The processivity of NS3 was not found to be significantly dependent on either [ATP] or [NS3] (Supplementary Fig. 6). However, the processivity of NS3 is strongly dependent on force¹⁶ (Fig. 4a): whereas an average of 18 bp are unwound at 5 pN before the enzyme dissociates, an average of 53 bp are unwound at 17 pN. In contrast, force does not affect the pause duration or stepping velocity of NS3 (Fig. 4b), indicating that NS3 might be limited not by strand separation but rather by translocation. Alternatively, the rate of strand separation might not be affected by force either because of protection of the RNA by NS3 or because the strand separation reaction coordinate in the presence of NS3 is orthogonal to the external force. Taken together, these results imply

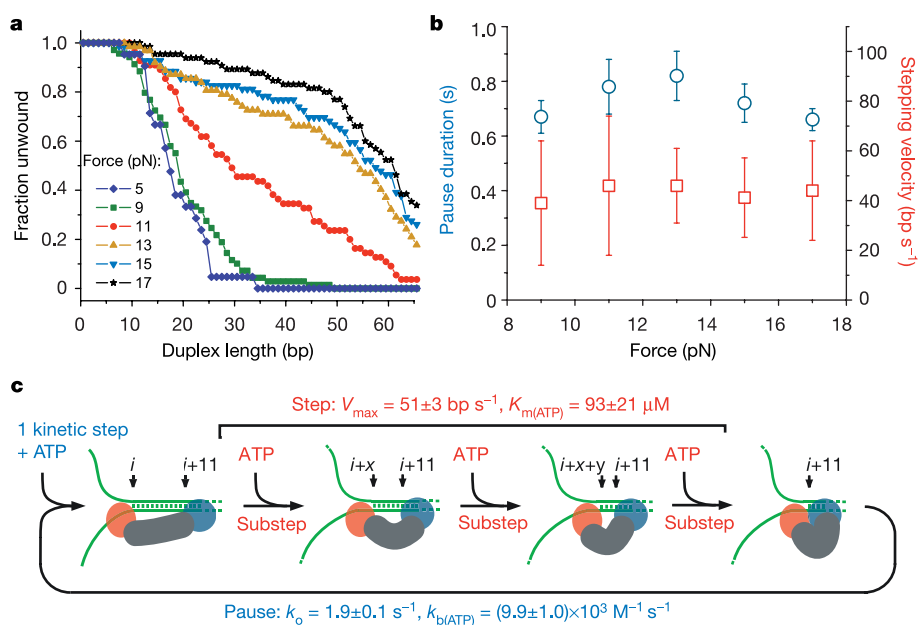


Figure 4 | Effect of force on the behaviour of NS3, and proposed model of action. **a**, Fraction of duplex RNA of different lengths that are unwound processively at different forces. The mean processivity of NS3 is 18, 20, 33, 46, 48 and 53 bp at 5, 9, 11, 13, 15 and 17 pN (21, 69, 64, 62, 69 and 67 traces, respectively; 1-bp bins). **b**, Stepping velocities (red squares; means $\pm$ s.d.) and pause duration (blue circles; means and 95% fit confidence intervals) in 5 nM NS3 at 9, 11, 13, 15 and 17 pN (18, 35, 19, 53 and 39 traces, respectively). Pause durations plotted here were determined from a single-exponential fit of the pause duration histograms (Supplementary Fig. 4). **c**, Proposed model of NS3 translocation and unwinding. The helix opener site (red ellipse) unwinds the substrate in substeps of 2–5 bp (x and y designate substep sizes) triggered by ATP ($V_{\max}$, K_m). The translocator (blue circle) contacts dsRNA every 11 bp, and possibly elsewhere during the cycle as well, which requires ATP binding (k_b) and an [ATP]-independent kinetic step (k_o).

that the applied force increases the average time that NS3 remains bound to its substrate (Fig. 4a, b). These results indicate a possible competition between strand reannealing and NS3 binding and that NS3 might be a processive single-strand translocase in the absence of competition from strand reannealing. Future experiments will probe to what extent NS3 uses a brownian ratchet mechanism and an active duplex-melting mechanism. The differential effects of force on NS3 velocity and processivity indicate that the forward movement and dissociation of NS3 are biochemically distinct events.

There are important similarities and differences between the NS3 activities reported here and those observed in previous bulk experiments. Kinetic features are similar: for example, the stepping velocity at saturating [ATP] is $35 \pm 4 \text{ bp s}^{-1}$ in bulk measurements and $51 \pm 3 \text{ bp s}^{-1}$ as measured here, and K_m is $160 \pm 6 \mu\text{M}$ in bulk measurements with the helicase domain of NS3 (ref. 22) whereas the apparent K_m measured here is $163 \pm 10 \mu\text{M}$ (Supplementary Discussion). However, previous measurements of step size ($18 \pm 2 \text{ bp}$)²⁰ are larger than reported here ($11 \pm 3 \text{ bp}$). This difference probably stems from the fact that different NS3 oligomeric states are monitored in the two types of experiment (Supplementary Discussion). During single-cycle bulk assays of RNA unwinding, the functional species studied is an NS3 dimer. This dimer requires long 3' single-stranded (ss) RNA overhangs (at least 14 nt) for initiation and displays a strictly concentration-dependent unwinding amplitude²⁰; moreover, it has a second-order rate constant for the kinetics of functional complex formation⁹ (V.S. and A.M.P., unpublished observations). In the present assay, the active species is an NS3 monomer that initiates unwinding within seconds at short 3' overhangs (4 nt or longer; Supplementary Fig. 7) and shows no concentration dependence of pause duration, stepping velocity or processivity (Supplementary Figs 6b and 8). In a different bulk assay we show that the NS3 monomer indeed has ATP-dependent helicase activity in the absence of force, and that it unwinds with processivity below that observed at 5 pN in the single-molecule assay (data not shown), consistent with force increasing the processivity of the NS3 monomer.

The similarities between the activities of an NS3 monomer and dimer provide a perspective on the role of helicase dimerization for NS3, and perhaps other helicases. Although multimeric helicase assemblies may benefit from an increased unwinding processivity²⁴, oligomerization does not itself seem to introduce new features in the unwinding mechanism. The NS3 monomer may be the engine of the dimer, whose processivity can be increased by assistance either from

another NS3 molecule (and perhaps even from a different protein species) or by external force. The basis for the observed difference in step size of an NS3 monomer and of a dimer is not currently well understood.

One possible model among others (Supplementary Discussion) to account for the observed 11-bp periodicity and its substep structure is shown in Fig. 4c, in which each NS3 monomer has two RNA-binding sites and in which ATP binding coordinates both translocation and unwinding in the fashion of an inchworm^{5,6} (Supplementary Discussion). We designate one site as the translocator and the other as the helix opener: for each mechanistic cycle, the translocator moves by an 11-bp step to contact double-stranded (ds) RNA ahead^{25,26} of the fork, and the helix opener moves by substeps of 2–5 bp. A lower rate of ATP binding during a pause ($(9.9 \pm 1.0) \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$; Supplementary Fig. 4) than before subsequent substeps (at least $5 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$, Supplementary Fig. 5a) indicates that these two ATP-binding events might be distinct and that ATP binding might be required both for translocator movement (blue ATP in Fig. 4c) and for strand separation by the helix opener (red ATP in Fig. 4c; Supplementary Discussion).

With unprecedented spatial and temporal resolution, the present strand displacement assay provides a powerful new method of investigating individual mechanistic cycles of many other enzymes that separate nucleic acid strands in the course of their cellular functions, such as ribosomes and polymerases.

METHODS

RNA substrate. The hairpin substrate sequences were cloned into the *EcoRI* and *HindIII* sites of pBR322 (NEB). To generate the hairpin DNA fragment, four oligonucleotides (65–100 nt each) were sequentially annealed. After cloning, double digestion of the circular construct by *NcoI* (designed at the hairpin loop) and *SylI* yielded two DNA fragments, each containing half of the hairpin sequence. The two fragments were gel purified and sequenced separately, confirming the full hairpin sequence. A transcription template (pBR322 bases 3821–628 with hairpin insert) was obtained by polymerase chain reaction (PCR) with a T7 promoter-appended primer, and transcribed *in vitro* (Ambion). The 1,287-nt RNA product was annealed to pBR322-based DNA handles (Fig. 1a); the 599-bp handle had a 5' biotin group (added through a PCR primer) and the 535-bp handle a 3' digoxigenin group (added with *exo*-Klenow fragment (NEB) after 3' recessed end-generation by *EcoRI*). The RNA sequence between the handles consisted of a 20-nt single-stranded spacer (shortened to 0 nt without effect), both arms of the 60-bp hairpin (underlined) separated by a tetraloop, and a 19-nt single-stranded spacer for NS3 loading (shortened to 4 nt without effect): 5'-UCUCAUGCAGGACAGUCGGAGGGAGACACUACGUUCGGACUAGUGUACUCUGACUUGAGACUACUGACAUCAGAUCCAGAUCCUCCCCAUGG

GAGAUCUGGAUGUCAGUAGUCUCAAGUCAGAGUACACUAGUCCGAAC GUAGUGCUCACAGGAGCUCAGCUAUCAGAA-3'.

Unless otherwise mentioned, the above sequence (RNA1) was used. The pairing, geometry and attachment of the substrate were confirmed (Supplementary Discussion). Substrates of different sequences (RNA2, RNA3 and RNA4; Supplementary Fig. 2) were synthesized as above. Substrates had two helicase-loading sites, namely the ssRNA regions located 3' to the hairpin and the 535-bp DNA/RNA handle. However, the extension change resulting from DNA/RNA unwinding was one-tenth of that for hairpin unwinding. Replacing the hairpin by a 29-nt ssRNA region resulted in no unwinding detection under standard conditions.

NS3 protein. Full-length NS3 from HCV genotype 1a was expressed (pQE40 plasmid, N-terminal His₆ tag) and purified with a protocol described elsewhere²⁷, with small modifications. [NS3] was measured with the Bradford assay. Although some earlier bulk measurements were done with NS3 genotype 1b²⁰, all bulk measurements stated here refer to NS3 genotype 1a unless otherwise mentioned.

Experiments with optical tweezers. We used a dual-beam optical tweezers instrument²⁸ to manipulate individual RNA molecules. Molecule characterization was performed as described elsewhere²⁹. The RNA was unfolded and refolded by moving (200 nm s⁻¹) the 2.2-μm streptavidin-coated bead (SpheroTech), connected to a piezoelectric stage through a micropipette, relative to the 2.9-μm anti-digoxigenin-coated bead³⁰, held in the laser trap. The distance between the folded state and the transition state²⁹ of the 60-bp hairpin was large (more than 10 nm), resulting in a narrow force range over which unfolding took place. Force-catalysed unfolding at 1.4 pN or more below the mean unfolding force was not observed. Buffer U consisted of 20 mM MOPS, 30 mM NaCl, 0.9% v/v glycerol, 0.75 mM MgCl₂, 0.1% Tween 20, 2 mM dithiothreitol, pH 6.5 at 22 °C. Data were collected at 60 Hz and the force detector signal was smoothed with an RC filter with a time constant of 14 ms. In constant-force feedback mode, force was constant to 0.1 pN. In the absence of NS3, two standard deviations in extension covered 1.4 nm at 60 Hz, which translates to a 2-bp resolution under our molecular geometry at 17 pN tension (Supplementary Discussion). Data were collected in the 10 min after injection of NS3 and ATP (Supplementary Discussion). Putative unwinding was confirmed by pulling the substrate after the constant-force period (Fig. 1b): a shortened mechanical unfolding transition or its absence indicated partial or full unwinding, respectively. For each RNA molecule attached between beads, we collected an average of ten unwinding traces.

Data analysis. The worm-like chain model¹⁷ was used with a ssRNA contour length of 0.59 nm per nucleotide and a persistence length of 1 nm (ref. 29). NS3 binding to ssRNA did not significantly change its contour and persistence lengths (Supplementary Discussion). The step size was the spatial frequency with the highest power (Fourier analysis) of a pairwise distance distribution from a single trace. An identical step size was obtained if the power spectra of individual pairwise distance distributions were summed (Supplementary Fig. 2f) and the highest power was identified. Steps were detected by scanning for maxima in the slope of the unwinding curve within a running-window (MATLAB custom-written programs). The window size was larger than step durations, and data points within maximal-slope windows were cropped on the basis of the mean and standard deviation of constant-extension regions. Pause durations were taken as the intervals between two successive steps and the velocity of a step was its best linear-fit slope (of the raw data). As [ATP] decreased, the number of data points per step increased and the best linear-fit R^2 decreased. The maximum velocity that the constant-force feedback could follow was 172 ± 17 nm s⁻¹, well above the maximum observed NS3 velocity. Substeps were required to belong to a previously detected step and were detected with a smaller running window than steps, whereas subpauses (periods between substeps within a step) were required to be longer than 80 ms.

Received 23 June; accepted 17 October 2005.

1. Jankowsky, E., Gross, C. H., Shuman, S. & Pyle, A. M. The DExH protein NPH-II is a processive and directional motor for unwinding RNA. *Nature* **403**, 447–451 (2000).
2. Lucius, A. L. & Lohman, T. M. Effects of temperature and ATP on the kinetic mechanism and kinetic step-size for *E. coli* RecBCD helicase-catalyzed DNA unwinding. *J. Mol. Biol.* **339**, 751–771 (2004).
3. Kolykhalov, A. A., Mihalik, K., Feinstein, S. M. & Rice, C. M. Hepatitis C virus-encoded enzymatic activities and conserved RNA elements in the 3' untranslated region are essential for virus replication *in vivo*. *J. Virol.* **74**, 2046–2051 (2000).
4. Frick, D. N. Helicases as antiviral drug targets. *Drug News Perspect.* **16**, 355–362 (2003).
5. Velankar, S. S., Soultanas, P., Dillingham, M. S., Subramanya, H. S. & Wigley, D. B.

Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* **97**, 75–84 (1999).

6. Bianco, P. R. & Kowalczykowski, S. C. Translocation step size and mechanism of the RecBC DNA helicase. *Nature* **405**, 368–372 (2000).
7. Delagoutte, E. & von Hippel, P. H. Helicase mechanisms and the coupling of helicases within macromolecular machines. Part II: Integration of helicases into cellular processes. *Q. Rev. Biophys.* **36**, 1–69 (2003).
8. Dimitrova, M., Imbert, I., Kiény, M. P. & Schuster, C. Protein-protein interactions between hepatitis C virus nonstructural proteins. *J. Virol.* **77**, 5401–5414 (2003).
9. Pang, P. S., Jankowsky, E., Planet, P. J. & Pyle, A. M. The hepatitis C viral NS3 protein is a processive DNA helicase with cofactor enhanced RNA unwinding. *EMBO J.* **21**, 1168–1176 (2002).
10. Levin, M. K., Gurjar, M. & Patel, S. S. A Brownian motor mechanism of translocation and strand separation by hepatitis C virus helicase. *Nature Struct. Mol. Biol.* **12**, 429–435 (2005).
11. Kim, J. L. *et al.* Hepatitis C virus NS3 RNA helicase domain with a bound oligonucleotide: the crystal structure provides insights into the mode of unwinding. *Structure* **6**, 89–100 (1998).
12. Dohoney, K. M. & Gelles, J. Chi-sequence recognition and DNA translocation by single RecBCD helicase/nuclease molecules. *Nature* **409**, 370–374 (2001).
13. Bianco, P. R. Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. *Nature* **409**, 374–378 (2001).
14. Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the *Escherichia coli* Rep helicase. *Nature* **419**, 638–641 (2002).
15. Perkins, T. T., Li, H. W., Dalal, R. V., Gelles, J. & Block, S. M. Forward and reverse motion of single RecBCD molecules on DNA. *Biophys. J.* **86**, 1640–1648 (2004).
16. Dessinges, M. N., Lionnet, T., Xi, X. G., Bensimon, D. & Croquette, V. Single-molecule assay reveals strand switching and enhanced processivity of UvrD. *Proc. Natl Acad. Sci. USA* **101**, 6439–6444 (2004).
17. Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599–1600 (1994).
18. Levin, M. K., Yang, Y. H. & Patel, S. S. The functional interaction of the hepatitis C virus helicase molecules is responsible for unwinding processivity. *J. Biol. Chem.* **279**, 26005–26012 (2004).
19. Singleton, M. R., Dillingham, M. S., Gaudier, M., Kowalczykowski, S. C. & Wigley, D. B. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature* **432**, 187–193 (2004).
20. Serebrov, V. S. & Pyle, A. M. Periodic cycles of RNA unwinding and pausing by hepatitis C virus NS3 helicase. *Nature* **430**, 476–480 (2004).
21. Ali, J. A. & Lohman, T. M. Kinetic measurement of the step size of DNA unwinding by *Escherichia coli* UvrD helicase. *Science* **276**, 377–380 (1997).
22. Levin, M. K. & Patel, S. S. Helicase from hepatitis C virus, energetics of DNA binding. *J. Biol. Chem.* **277**, 29377–29385 (2002).
23. Levin, M. K., Gurjar, M. M. & Patel, S. S. ATP binding modulates the nucleic acid affinity of hepatitis C virus helicase. *J. Biol. Chem.* **278**, 23311–23316 (2003).
24. Tackett, A. J., Chen, Y., Cameron, C. E. & Raney, K. D. Multiple full-length NS3 molecules are required for optimal unwinding of oligonucleotide DNA *in vitro*. *J. Biol. Chem.* **280**, 10797–10806 (2005).
25. Tackett, A. J., Wei, L., Cameron, C. E. & Raney, K. D. Unwinding of nucleic acids by HCV NS3 helicase is sensitive to the structure of the duplex. *Nucleic Acids Res.* **29**, 565–572 (2001).
26. Bianco, P. R. Hepatitis C NS3 helicase unwinds RNA in leaps and bounds. *Lancet* **364**, 1385–1387 (2004).
27. Yao, N. *et al.* Structure of the hepatitis C virus RNA helicase domain. *Nature Struct. Biol.* **4**, 463–467 (1997).
28. Smith, S. B., Cui, Y. & Bustamante, C. Optical-trap force transducer that operates by direct measurement of light momentum. *Methods Enzymol.* **361**, 134–162 (2003).
29. Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I. Jr & Bustamante, C. Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**, 733–737 (2001).
30. Bryant, Z. *et al.* Structural transitions and elasticity from torque measurements on DNA. *Nature* **424**, 338–341 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. V. Le from the Schering-Plough Research Institute for the NS3 plasmid; S. B. Smith, P. T. X. Li, Y. R. Chemla and J.-C. Liao for discussions and technical help; T. M. Lohman for critical reading of the manuscript, and members of our laboratories for discussions and critical reading of the manuscript. This research was supported by CIHR and FQRNT doctoral fellowships (S.D.), an NIH postdoctoral fellowship (R.K.B.), NIH (I.T., A.M.P., C.B.), DOE (C.B.), and HHMI grants to investigators A.M.P. and C.B.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.B. (carlos@alice.berkeley.edu).

Functional waters in intraprotein proton transfer monitored by FTIR difference spectroscopy

Florian Garczarek¹ & Klaus Gerwert¹

Much progress has been made in our understanding of water molecule reactions on surfaces¹, proton solvation in gas-phase water clusters^{2,3} and proton transfer through liquids⁴. Compared with our advanced understanding of these physico-chemical systems, much less is known about individual water molecules and their cooperative behaviour in heterogeneous proteins during enzymatic reactions. Here we use time-resolved Fourier transform infrared⁵ spectroscopy (trFTIR) and *in situ* H₂¹⁸O/H₂¹⁶O exchange FTIR to determine how the membrane protein bacteriorhodopsin⁶ uses the interplay among strongly hydrogen-bonded water molecules, a water molecule with a dangling hydroxyl group and a protonated water cluster⁷ to transfer protons. The precise arrangement of water molecules in the protein matrix results in a controlled Grotthuss proton transfer, in contrast to the random proton migration that occurs in liquid water. Our findings support the emerging paradigm that intraprotein water molecules are as essential for biological functions as amino acids.

Ever since Grotthuss⁸ proposed his famous concept of fast proton transfer along water chains about 200 yr ago, physical chemists have been trying to understand exactly how protons migrate in liquid water. Today, the results from high-performance computer simulations show that the interconversion between an Eigen cation⁹ (H⁺·(H₂O)₄) and a Zundel cation¹⁰ (H⁺·(H₂O)₂) leads to proton migration in water⁴. In the Eigen complex the hydronium core (H₃O⁺) is bound to three H₂O molecules ((H₃O⁺)·(H₂O)₃), whereas in the Zundel cation the proton fluctuates between two water molecules (H₂O···H⁺···OH₂). In contrast to this random proton migration in liquid water, a more controlled proton transfer is crucial to many enzymatic reactions in proteins. Internal water-filled cavities have been discovered in several high-resolution structural models of membrane proteins but their functional role during proton translocation is not yet clear.

An excellent model system in which to study the role of protein internal water molecules is the well-characterized light-driven proton pump bacteriorhodopsin⁶, a seven α -helical membrane protein in which the chromophore retinal is bound by a protonated Schiff base. The absorption of light induces an all-*trans* to 13-*cis* isomerization of retinal and leads, after a few picoseconds, to a metastable protein state with an excess enthalpy of about 12 kcal mol⁻¹ (ref. 11) termed the 'K intermediate' (Fig. 1a). The first proton transfer takes place after ~80 μ s during the transition from the L to M intermediate. A proton transfers from the protonated Schiff base PSB to its counterion D85 at roughly the same time that a proton is released to the extracellular medium from the proton release group located near E204 and E194 (Fig. 1b). D96, located on the proton uptake side, reprotonates the Schiff base in the M to N transition¹² and is itself reprotonated from the cytoplasmic bulk. The proton release group receives a proton from D85 in the back reaction to the ground state. The net reaction is to pump a proton across the membrane.

Structural models of the bacteriorhodopsin ground state at 1.55 Å resolution have identified several internal fixed water molecules¹³. These models are deduced from X-ray diffraction data from crystallized proteins taken at 100 K and therefore represent only a static assembly. To characterize the dynamics of the water molecules in the protein, we simulated bacteriorhodopsin under physiological conditions, namely, at room temperature, in a lipid bilayer and completely surrounded by external water molecules¹⁴. Two extensive water densities at the extracellular side of bacteriorhodopsin were identified (Fig. 1c). The protonation state and connecting hydrogen bonds of these water molecules are crucial for the properties of internal water clusters in proteins.

We used *in situ* H₂¹⁶O/H₂¹⁸O exchange (Fig. 2a) and time-resolved FTIR difference spectroscopy (Fig. 2b–d) at room temperature to monitor the hydrogen-bonding state and protonation changes of internal water clusters during the bacteriorhodopsin photocycle. Figure 1d shows a typical three-dimensional representation of a time-resolved FTIR difference (relative to the ground state) data set for the bacteriorhodopsin photocycle. In such a difference spectrum, the positive bands originate from the formation of intermediates and the negative bands refer to the ground state. The appearance of the positive carbonyl band at 1,762 cm⁻¹ indicates transient protonation of D85 in the M intermediate. The much broader, negative, 'continuum absorbance' indicates a protonated water cluster, which deprotonates in the M intermediate¹⁵. The nature of the proton release group has been debated for many years^{15,16}, but has now been shown to be a protonated water cluster that is hydrogen-bonded to six side chains, including R82, E194 and E204 and three backbone groups⁷.

In addition to this lower water cluster at the release site, three water molecules have been found in the vicinity of the Schiff base (Fig. 1c, green). Figure 3a shows how these water molecules, together with each carboxylate oxygen of D85 and D212, might constitute a pentagonal structure^{17,18} that is hydrogen-bonded to the proton at the Schiff base. Such a pentagonal arrangement is a typical low-energy conformation in water known as a 'cyclic pentamer'¹⁹. Although the positions of the oxygen atoms (red spheres) of the water molecules can be determined from X-ray experiments, their orientations remain unknown.

In the conformation proposed in Fig. 3a, the hydroxyl group of water W401 does not participate in hydrogen bonding. Such dangling groups are also observed on aqueous surfaces¹. The hydroxyl stretch vibration of a dangling group shows a very sharp band at frequencies higher (>3,600 cm⁻¹) than observed in liquid water (3,500–3,200 cm⁻¹) owing to the missing hydrogen bond. A free hydroxyl stretch vibration exists in the bacteriorhodopsin ground state (see ref. 17 and citations therein) at 3,644 cm⁻¹ (Fig. 2a). This band shifts from 3,644 to 3,633 cm⁻¹ during *in situ* H₂¹⁶O/H₂¹⁸O exchange measurement at room temperature and

¹Lehrstuhl für Biophysik, Ruhr-Universität Bochum, D-44780 Bochum, Germany.

therefore clearly belongs to a dangling water group. The water must be located inside the protein and its exact position can be determined with mutants. The $3,644/3,633\text{ cm}^{-1}$ difference band disappears when an asparagine residue (N) replaces the aspartate at position 85 in the D85N mutant. As this mutant alters the hydrogen bonding of the pentagonal arrangement, the dangling group water must belong to the upper water cluster (Fig 1, green) and is most probably W401 as shown in Fig. 3a. A dangling group in the upper water cluster at room temperature is evidence that the water molecules in this cluster are immobilized even under physiological conditions. This is notable, because it has been proposed by quantum chemical calculations that the presence of W401 at this position is responsible for the ionic state of D85 (ref. 18).

The water arrangement undergoes a considerable change after photoexcitation²⁰. The dangling hydroxyl group of W401 disappears in the M intermediate (Fig. 2a, green), probably owing to its higher mobility in M as proposed by molecular dynamic simulations²¹. The former free hydroxyl group becomes hydrogen-bonded and the corresponding hydroxyl stretch shifts to lower wavenumbers between $3,600$ and $3,200\text{ cm}^{-1}$, where the overlap with the large background absorbance of the solvent water makes it impossible to resolve. Notably, a new water dangling group absorbing at $3,671/3,660\text{ cm}^{-1}$ appears. It represents a water at the cytoplasmic side¹⁷ that is involved in the reprotonation of the Schiff base by intruding water molecules^{6,22}.

As compared with free hydroxyl groups, the stretching band of hydrogen-bonded hydroxyl groups shows a pronounced shift to lower wavenumbers (red shift) and a strong broadening. Figure 2b shows the time-resolved FTIR difference spectrum between $3,050$ and $1,720\text{ cm}^{-1}$ taken $100\text{--}200\text{ ns}$ and $3\text{--}10\text{ }\mu\text{s}$ after laser excitation. It represents the changes in the K and L intermediates, respectively. The

black baseline represents the absorbance changes caused by the actinic laser flash, which thermally heats water in the hydrated protein sample²³. Relative to this black line, there is an additional broad negative band between $3,000$ and $2,400\text{ cm}^{-1}$ in both intermediates. A similar absorbance change is observed in low-temperature spectra of the K intermediate of bacteriorhodopsin and it has been proposed that this band consists predominantly of absorptions of strong hydrogen-bonded waters superposed by the N–H stretch of the strong hydrogen-bonded Schiff base^{17,24}.

Mutations of D85 and D212 affect the broad water absorbance band, but mutations of residues hydrogen-bonded to the lower clusters S193A, E194Q and E204Q show no effect (Fig. 2b). Hydrogen bonds involving ions are generally stronger than those between water molecules. Because W402 is hydrogen-bonded to the positively charged Schiff base and to the two negatively charged carboxylates, D85 and D212 (see Fig. 3a), it is the most likely candidate responsible for this broad absorbance. The disappearance of this absorption in K and L shows that the retinal isomerization in the primary light reaction weakens the strong hydrogen bonds and that they are not re-established in the L intermediate. This result, obtained under physiological conditions, contrasts with the low-temperature measurements, which show reduced amplitude in L. The low-temperature results led to a proposal of a hydration switch model²⁵, which obviously does not describe the situation at room temperature.

The hydroxyl stretch bands seem to shift to and get lost in the absorbance region of the solvent liquid water between $3,600$ and $3,200\text{ cm}^{-1}$. An approximate relationship between a hydroxyl stretch band shift and bond enthalpies²⁶ yields $-\Delta H = 0.31(\Delta\nu)^{1/2}$, where ΔH is in kcal mol^{-1} and $\Delta\nu$ is in cm^{-1} . The frequency shift of one hydroxyl group from $\sim 2,800\text{ cm}^{-1}$ to $\sim 3,400\text{ cm}^{-1}$, that is, from a strong to an average hydrogen bond, indicates an enthalpy loss of

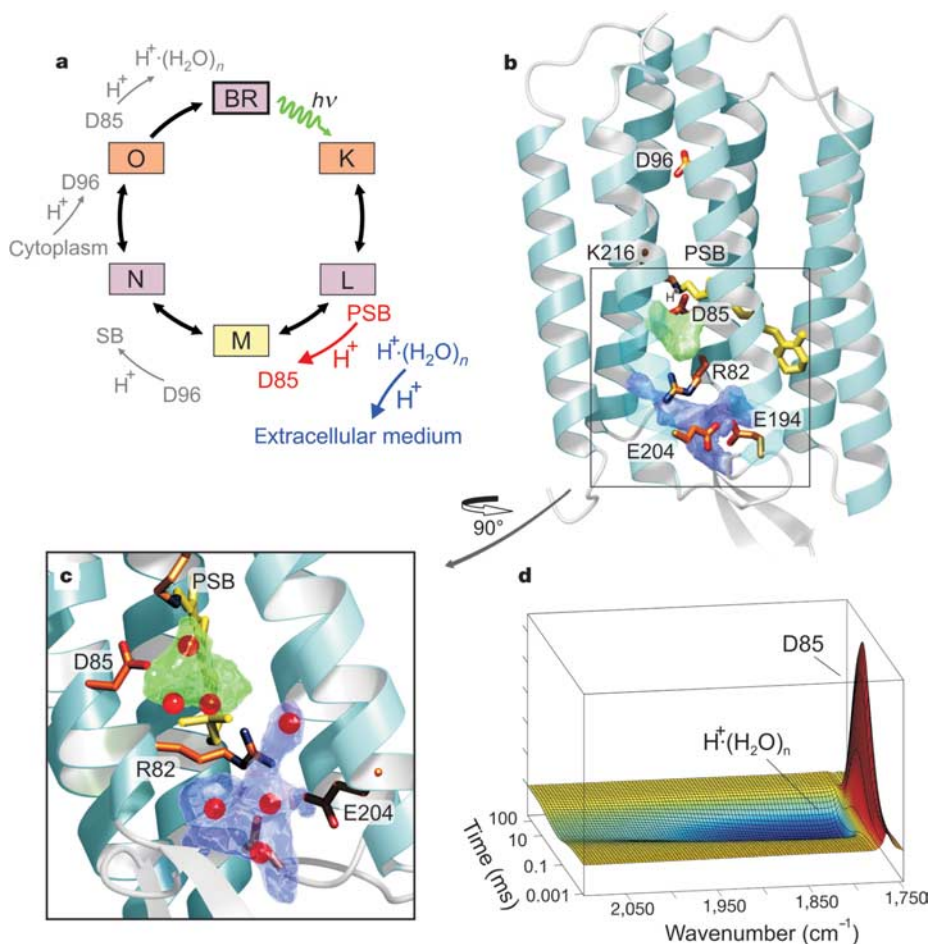


Figure 1 | Structure and proton transfer steps of bacteriorhodopsin. **a**, Photocycle scheme showing the main steps in light-driven proton transport. BR, ground state bacteriorhodopsin; PSB, protonated Schiff base. **b**, Structural model of bacteriorhodopsin (PDB code 1C3W; ref. 13) showing its key residues, including the protonated Schiff base by which the chromophore retinal (yellow) is bound to the protein at K216, as well as D85, R82, E194, E204 and D96. The extracellular side shows the space occupied by fluctuating internal water molecules in green (upper water cluster) and blue (lower water cluster)¹⁴. **c**, Rotation by 90° of the boxed area in **b** to show the position of the water oxygen atoms (red spheres) in the water densities, as determined by X-ray crystallography¹³. The two front helices have been omitted for clarity. **d**, Time-resolved absorbance changes during the photocycle between $2,100$ and $1,750\text{ cm}^{-1}$. At $1,762\text{ cm}^{-1}$ protonation of D85 (red), and between $2,100$ and $1,800\text{ cm}^{-1}$ deprotonation of a water complex (blue), can be followed.

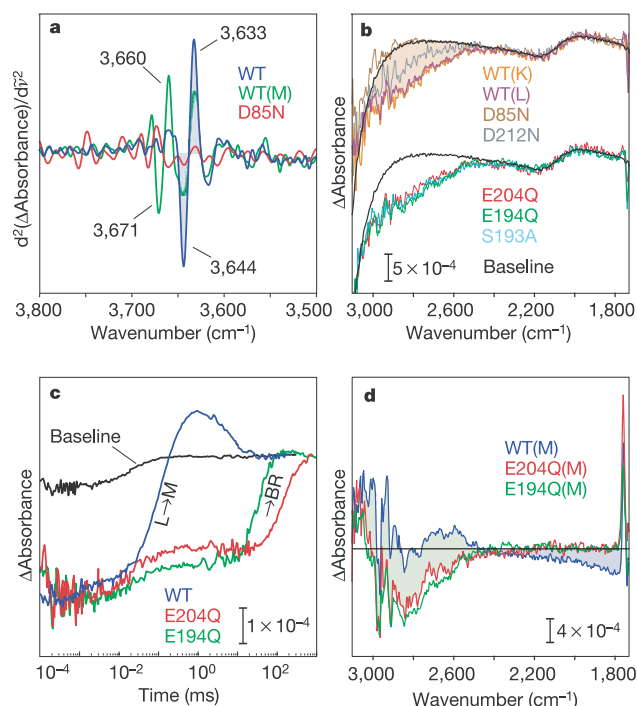


Figure 2 | FTIR measurements of the internal water molecules of bacteriorhodopsin. **a**, *In situ* $\text{H}_2^{16}\text{O}/\text{H}_2^{18}\text{O}$ exchange measurement for wild type (WT) in the ground state (blue), D85N in the ground state (red) and the wild-type M intermediate (intermediate accumulation $\approx 60\%$; green). The vibration of a water dangling hydroxyl group can be identified by its isotopic shift from $3,644$ to $3,633\text{ cm}^{-1}$ in the wild-type ground state. This hydroxyl group is absent in both D85N and the wild-type M intermediate. In the latter, a new difference band appears at $3,671/3,660\text{ cm}^{-1}$. **b**, Time-resolved FTIR data in the $3,050\text{--}1,720\text{ cm}^{-1}$ region showing the difference spectra for the K and L intermediates of wild type and for the indicated mutants. The black baseline is not flat owing to the actinic laser flash²³. The broad negative band (highlighted in red) indicates the breaking of strong hydrogen-bonded water molecules. **c**, Time course of the absorbance changes integrated between $2,900$ and $2,600\text{ cm}^{-1}$ for wild type and two mutants, indicating the breaking ($\Delta\text{Absorbance} < 0$) and merging ($\Delta\text{Absorbance} \geq 0$) of strong hydrogen bonds. BR, ground state bacteriorhodopsin. **d**, As **b**, but for the M intermediate. The absorbance of regenerated strong hydrogen-bonded water is highlighted in green and the continuum absorbance change is highlighted in blue.

about 5.6 kcal mol^{-1} . Thus, it needs just the weakening of two strong hydrogen bonds to store the K intermediate enthalpy of about 12 kcal mol^{-1} (ref. 11). The upper water cluster may serve as a temporary energy storage vehicle in the primary photo activation²⁴. The enthalpy lost in the primary reaction by the breaking of strong hydrogen bonds could probably be used later during the photocycle for proton transfer.

Figure 2c shows the time dependence of the strong hydrogen-bonded region. The black line again represents the time course of the artificial heat signal. The instantaneous disappearance of the wild-type trace in blue indicates the above-mentioned cleavage of the strong hydrogen bonds in the upper water cluster. Notably, new strong hydrogen bonds form in the L to M transition ($\sim 80\text{ }\mu\text{s}$), concomitant with the deprotonation of the water complex at the release site¹⁵. In E204Q and E194Q mutants of bacteriorhodopsin, by contrast, these strong hydrogen bonds do not form in M, but instead form at the end of the photocycle with a delay in proton release (by a factor of $\sim 10^3$) (refs 6, 16). The strong hydrogen bonds in M might form between W404 and the negatively charged E204 and E194 (Fig. 3b). As expected for such locations, the difference spectra of E204Q and E194Q show deviations as compared with the wild-type protein in the strong hydrogen-bonded region (Fig. 2d, green). The formation of new strong hydrogen bonds among E204, E194 and W404 may provide the enthalpy for release of the proton to the bulk.

The released proton is formally stored in a water cluster (Fig. 3a, blue). Whether the excess proton is stored in a Zundel-like or Eigen-like form remains open. Because the experimentally observed spectra of bacteriorhodopsin agree well with quantum chemical calculations of the spectral dependence of a Zundel cation²⁷, but not an Eigen cation, we previously proposed a Zundel-like protonated water complex for bacteriorhodopsin⁷. The calculations give a maximum at $\sim 2,500\text{ cm}^{-1}$ for the Eigen cation and at $\sim 1,800\text{ cm}^{-1}$ for the Zundel cation, the latter being in agreement with observation. However, mid infrared spectra have now been taken of the protonated water clusters $\text{H}^+(\text{H}_2\text{O})_n$ with $2 \leq n \leq 11$ water molecules in the gas phase³. Compared with the sharp absorbance bands of the protonated gas-phase water clusters, the bands in bacteriorhodopsin are much broader owing to the coupling with the thermal fluctuating protein environment. The gas-phase measurements show that, in addition to protonated water clusters with $n = 2$ (Zundel), those with $n = 3$ and $n = 5$ also primarily absorb around $1,800\text{ cm}^{-1}$, and

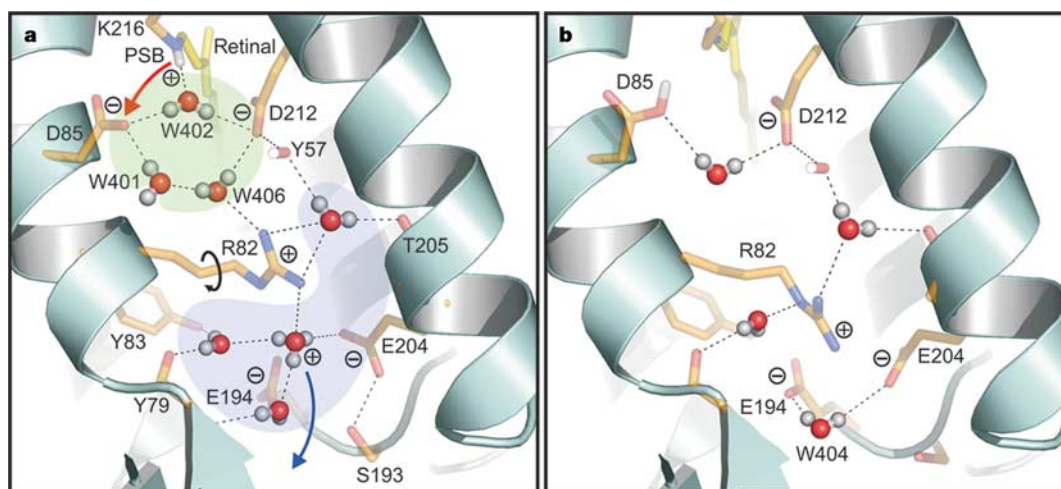


Figure 3 | Proton transfer via protein internal bound water molecules in the extracellular region of bacteriorhodopsin as derived from the FTIR data. **a**, Annotation of the area shown in Fig. 1c with water molecules including the hydrogens and catalytic residues (based on PDB 1C3W; ref. 13). Protonation of D85 induces R82 movement. **b**, Structure of the M

intermediate (based on PDB 1C8S; ref. 29). Release of the excess proton is caused by the movement of R82. The strongly hydrogen-bonded water 402, the water molecule with a dangling hydroxyl group (W401) and the protonated water complex are no longer observed in M.

therefore such clusters may also underlie the observed spectral changes in bacteriorhodopsin. The $n = 3$ cluster forms a hydronium ion (H_3O^+) that, unlike the classical Eigen cluster, no longer benefits from the symmetrical solvation of three water molecules. An asymmetrically solvated H_3O^+ ion is thus as likely as a Zundel cation to be the core of the protonated water cluster of bacteriorhodopsin. At present, we cannot distinguish between these two possibilities or the possibility that the proton fluctuates between both forms at the release site.

Integrating the data from years of experimentation on bacteriorhodopsin⁶, we can propose the following molecular mechanism for proton release. The light induced retinal isomerization breaks the interconnections in the water–aspartate pentamer (Fig. 3a, green), which stores parts of the excitation enthalpy owing to hydrogen-bond cleavage and reduces the pK of the central proton binding site. The latter results in the transfer of a proton from the protonated Schiff Base to D85, which induces a movement of R82 (refs 28, 29, and Fig. 3b). R82 contributes to solvation of the lower protonated water cluster^{7,14}. In bacteriorhodopsin, amino acids have the role of the so-called ‘second shell waters’ that solvate protonated waters in liquids. In liquids, the thermal fluctuation-driven release of a water from the second hydration shell, rather than the fast fluctuation of the proton, is rate limiting. With amino acids replacing these waters, the arbitrary release of a second shell water is no longer rate limiting; rather, the protein-controlled R82 movement towards the protonated water cluster determines the proton release owing to electrostatic interactions (Fig. 3b). The pK of the protonated cluster decreases from 9.5 in the ground state to below 5.3 in the M intermediate. The enthalpy for this process might be gained by the formation of new strong hydrogen bonds, probably among W404, E204, E194 and R82 (Fig. 3b).

Similar to the way in which they provide proton transfer in bacteriorhodopsin, single water molecules may transfer protons at the proton uptake side in photosynthetic proteins and also in redox-driven proton pumps such as cytochrome *c* oxidase.

METHODS

Mutagenesis and mutant expression. Purple membrane sheets containing bacteriorhodopsin were isolated and the site-specific mutants were prepared as described⁷.

Time-resolved measurements. Time-resolved measurements were made by the step-scan FTIR technique⁷. Purple membranes were suspended (200 µg in 1 M KCl and 100 mM Tris-HCl buffer; pH 7), centrifuged (2 h at 200,000g) and squeezed between two CaF_2 windows. Spectra were taken with a spectral resolution of 4 cm^{-1} and a time resolution of 30 ns.

In situ $\text{H}_2^{16}\text{O}/\text{H}_2^{18}\text{O}$ exchange measurements. Pure membrane sheets were suspended at a concentration of 5 mg ml^{-1} in 0.5 mM KCl and 0.5 mM Tris-HCl buffer (pH 8). Two 10-µl aliquots of this solution were dried on two CaF_2 windows. The experimental procedure was similar to *in situ* H/D exchange measurements⁷ except that here isotopically labelled H_2^{18}O replaced deuterium dioxide and the measurements were taken at room temperature for the ground state (238 K for the M intermediate). Because of this higher temperature, a larger amount of water vapour was present in the sample. To separate the water vapour bands from the bands of interest, we subtracted a protein-free reference spectrum from the experimental data. To depress the large and broad water absorbance bands of solvent water, the second derivatives of the exchange difference spectra were calculated as described⁷ (in brief, $\frac{d^2}{d\nu^2}\text{H}_2^{16}\text{O}/\text{H}_2^{18}\text{O}$ spectrum).

Figures. Figures 1b, c and 3 were prepared with the programs MOLMOL³⁰, POV-Ray (<http://www.povray.org>) and PyMOL (<http://www.pymol.org>).

Received 9 August; accepted 8 September 2005.

Published online 9 November 2005.

- Marx, D. Throwing tetrahedral dice. *Science* **303**, 634–636 (2004).
- Zwier, T. S. The structure of protonated water clusters. *Science* **304**, 1119–1120 (2004).
- Headrick, J. M. *et al.* Spectral signatures of hydrated proton vibrations in water clusters. *Science* **308**, 1765–1769 (2005).
- Marx, D., Tuckerman, M. E., Hutter, J. & Parrinello, M. The nature of the hydrated excess proton in water. *Nature* **397**, 601–604 (1999).

- Kötting, C. & Gerwert, K. Proteins in action monitored by time-resolved FTIR spectroscopy. *Chem. Phys. Chem.* **6**, 881–888 (2005).
- Lanyi, J. K. Bacteriorhodopsin. *Annu. Rev. Physiol.* **66**, 665–688 (2004).
- Garczarek, F., Brown, L. S., Lanyi, J. K. & Gerwert, K. Proton binding within a membrane protein by a protonated water cluster. *Proc. Natl Acad. Sci. USA* **102**, 3633–3638 (2005).
- de Grothhuss, C. J. T. Sur la décomposition de l'eau et des corps qu'elle tient en dissolution à l'aide de l'électricité galvanique. *Ann. Chim.* **58**, 54–74 (1806).
- Eigen, M. Proton transfer, acid-base catalysis, and enzymatic hydrolysis. Part I: Elementary processes. *Angew. Chem. Int. Edn Engl.* **3**, 1–19 (1964).
- Zundel, G. in *The Hydrogen Bond—Recent Developments in Theory and Experiments* (ed. Sandorfy, C.) 683–766 (Nort-Holland, Amsterdam, 1976).
- Birge, R. R. *et al.* Revised assignment of energy storage in the primary photochemical event in bacteriorhodopsin. *J. Am. Chem. Soc.* **113**, 4327–4328 (1991).
- Gerwert, K., Hess, B., Soppa, J. & Oesterhelt, D. Role of aspartate-96 in proton translocation by bacteriorhodopsin. *Proc. Natl Acad. Sci. USA* **86**, 4943–4947 (1989).
- Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291**, 899–911 (1999).
- Kandt, C., Schlitter, J. & Gerwert, K. Dynamics of water molecules in the bacteriorhodopsin trimer in explicit lipid/water environment. *Biophys. J.* **86**, 705–717 (2004).
- Rammelsberg, R., Huhn, G., Lubben, M. & Gerwert, K. Bacteriorhodopsins intramolecular proton-release pathway consists of a hydrogen-bonded network. *Biochemistry* **37**, 5001–5009 (1998).
- Dioumaev, A. K. *et al.* Existence of a proton transfer chain in bacteriorhodopsin: participation of Glu-194 in the release of protons to the extracellular surface. *Biochemistry* **37**, 2496–2506 (1998).
- Shibata, M. & Kandori, H. FTIR studies of internal water molecules in the Schiff base region of bacteriorhodopsin. *Biochemistry* **44**, 7406–7413 (2005).
- Hayashi, S. & Ohmine, I. Proton transfer in bacteriorhodopsin: Structure, excitation, IR spectra, and potential energy surface analyses by an *ab initio* QM/MM method. *J. Phys. Chem. B* **104**, 10678–10691 (2000).
- Liu, K., Brown, M. G., Cruzan, J. D. & Saykally, R. J. Vibration-rotation tunneling spectra of the water pentamer: Structure and dynamics. *Science* **271**, 62–64 (1996).
- Dencher, N. A., Sass, H. J., Buldt, G. Water and bacteriorhodopsin: structure, dynamics, and function. *Biochim. Biophys. Acta* **1460**, 192–203 (2000).
- Grudin, S., Buldt, G., Gordeliy, V. & Baumgaertner, A. Water molecules and hydrogen-bonded networks in bacteriorhodopsin—molecular dynamics simulations of the ground state and the M intermediate. *Biophys. J.* **88**, 3252–3261 (2005).
- Le Coutre, J., Tittor, J., Oesterhelt, D. & Gerwert, K. Experimental evidence for hydrogen-bonded network proton transfer in bacteriorhodopsin shown by Fourier-transform infrared spectroscopy using azide as catalyst. *Proc. Natl Acad. Sci. USA* **92**, 4962–4966 (1995).
- Garczarek, F., Wang, J., El-Sayed, M. A. & Gerwert, K. The assignment of the different infrared continuum absorbance changes observed in the 3000–1800 cm^{-1} region during the bacteriorhodopsin photocycle. *Biophys. J.* **87**, 2676–2682 (2004).
- Hayashi, S., Tajkhorshid, E., Kandori, H. & Schulten, K. Role of hydrogen-bond network in energy storage of bacteriorhodopsin's light-driven proton pump revealed by *ab initio* normal-mode analysis. *J. Am. Chem. Soc.* **126**, 10516–10517 (2004).
- Tanimoto, T., Furutani, Y. & Kandori, H. Structural changes of water in the Schiff base region of bacteriorhodopsin: proposal of a hydration switch models. *Biochemistry* **42**, 2300–2306 (2003).
- Rozenberg, M., Loewenschuss, A. & Marcus, Y. An empirical correlation between stretching vibration redshift and hydrogen bond length. *Phys. Chem. Chem. Phys.* **2**, 2699–2702 (2000).
- Rousseau, R., Kleinschmidt, V., Schmitt, U. W. & Marx, D. Unravelling water network protonation patterns in bacteriorhodopsin by theoretical IR spectroscopy. *Angew. Chem. Int. Edn Engl.* **43**, 4804–4807 (2004).
- Spassov, V. Z., Luecke, H., Gerwert, K. & Bashford, D. pK_a calculations suggest storage of an excess proton in a hydrogen-bonded water network in bacteriorhodopsin. *J. Mol. Biol.* **312**, 203–219 (2001).
- Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structural changes in bacteriorhodopsin during ion transport at 2 Å resolution. *Science* **286**, 255–260 (1999).
- Koradi, R., Billeter, M. & Wüthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55 (1996).

Acknowledgements We thank A. Hartz for protein preparation; E. Hofmann and C. Kandt for help with figure preparation; and R. S. Goody and D. Rumschitzki for help with English. This work was funded by the Deutsche Forschungsgemeinschaft.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.G. (gerwert@bph.uni-bochum.de).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

A plan for the future

Six days, three countries and far too many miles. Not an ambitious holiday schedule but a snapshot of my travels last November, as I rounded off a busy year of meeting young scientists around the world. In just over a week I met Max Planck PhD students at their annual meeting in Göttingen, Germany; joined young scientists interested in industry at the European Biotech Crossroads meeting in Lille, France; and sat in on an informal roundtable for students and postdocs at the University of Leeds, UK. Three countries, three completely different groups — and yet all echoed concerns I'd been hearing all year.

The students were aware that the odds of securing a tenure-track position in academia are low, but they didn't know the average success rate. I always say that on average only one in three is likely to land a long-term position. And all the students were vaguely aware that there are options 'outside' academia, but few knew exactly what they were, or how to take advantage of them.

My message to all these groups is the same: research your options with the diligence you spend on your science. Take an inventory of the skills you use in science — whether it is

interacting with intellectual-property people, writing papers or managing interdisciplinary collaborations. See what you enjoy and are good at. Then consider where else you could apply these skills. Cast a wide net. Be prepared to leave your home country. Don't discount government and industry jobs. And find creative ways to gain new skills — whether it be informal mentoring, short courses or yet more education.

As well as talking to young scientists about navigating their career, I have been trying to draw up some guidance to help them. One result is a collection of 'off the bench' career stories now gathered in one location under Editor's Choice on the *Naturejobs* homepage (www.nature.com/naturejobs). I hope that this collection will help scientists at all stages of their career realize their options and make informed choices about their next move.



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents: Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: r.asami@naturejpn.com

Gathering of the clans

Get in touch with your past.

Reinaldo José Lopes

You could see the pavilions for miles around in the bright summer morning. Only a little less conspicuous was the line of people moving slowly towards two signposts. "Got your marker ready? This way, please," said one. "First time? Come and shed your blood," informed the other. And above them, a bigger signpost shouted in fake-Celtic letters: "Welcome to the 5th GATHERING OF THE CLANS!"

The enthusiasm was almost palpable — or was it the smell of sweat? — except for a tiny segment of the line where a group of friends was arguing. They seemed to be having a hard time convincing one of them that yes, despite the evidence, this was going to be cool.

"Don't be such a baby, Pat. It'll be fun."

"And I still say it's gonna be ludicrous. And I hate needles. Do you really think they'll use a different one for every single person in this crowd?"

"You sound like a sissy. It's just a drop, for goodness sake! You know how efficient those sequencers are nowadays. They read the bases, tell you who was your great-great-grandpa of a couple of thousand years ago and that's it — you're free to drink mead and get the chicks."

"Yeah, Pat, what's wrong with that? After all, we're just connecting with our past, buddy. Thought you appreciated that."

"Look, I'm as likely to engage in hero-worship as anyone. It's the impersonality of it that bothers me. There was certainly a point in us claiming descent from Hengist and Horsa. Those guys at least had a story — they cut Finn and his gang to pieces and conquered Britain. Been there, done that. People can connect with that kind of stuff. But now you're asking me to worship a DNA sequence. No sir — I'd rather go drink mead with old Hengist in Valhalla."

Soon they were right in front of the gates, where a stout fellow in white was grinning at Pat — everyone, of course, had made sure he was the first to get in.

"Alright, mate, what's it gonna be? Mt? Y? Autosomal markers?"

Pat sighed. "Whatever. Surprise me."

Very carefully, his thumb was pierced. Ten seconds later, a robotic voice that seemed to be suffering from a dreadful case of personality emulation (of the irri-



tatingly happy sort) announced: "Congratulations! Your mtDNA has been assigned to the V haplogroup, fairly common in western and northern Europe, where it originated right after the Last Glacial Maximum! Your people were probably among the earliest and greatest artists the human race has known, creating those fine murals of extinct megafauna in Altamira and Lascaux. Way to go!"

"Thrilling," growled Pat. "Can I go now?"

"Hang on a sec," said the gatekeeper, "you need your totem!" He gave Pat a little plastic horse that looked like a poor imitation of the ones from Lascaux. Pat sighed still louder, grabbed the horse and moved on.

"Hey, what's wrong with Marvin there?" asked the gatekeeper.

"Oh, the usual thing. Don't talk to him about life," answered his friends, laughing.

The Gathering seemed to confirm Pat's worst fears. In one corner, somebody dressed as a Mongol warrior was calling "all Star-Cluster kids under ten" to learn how to shoot with a composite bow, just like Grandpa Genghis. A few yards away, some French families were being instructed in the minutiae of cannibalism among the Tupinambá tribes of Brazil — it turned out a young chief from that nation had married the daughter of a

Norman trader in the sixteenth century. Elsewhere, a rabbi was always ready in case you found out your chromosomes were Jewish and wanted an impromptu bar mitzvah.

Pat wandered miserably until he spotted a girl with long black hair who also seemed to be walking alone. Predictably, he was happy to inflict on her (Vera was the name, and she was from the Basque Country in Spain) all the talk about how ridiculous the whole Gathering concept was. Vera seemed to dig his grumpy-old-man charm, but didn't quite agree.

"I think everyone is aware of that stuff," she said. "But think of it for a second. Isn't it wonderful that all these different people are learning about a past that seems plain legend but is written in our blood?"

Besides, look at the scale of it. We used to think in terms of two or three generations at most. Now you can look back thousands of years and still recognize yourself."

Pat was still unconvinced. "I see what you mean. But I don't know quite how to feel about it. You see, I... Whoa!"

They had wandered to the very heart of the Gathering and were right in front of an awesome panel. Picture the largest family tree you have ever seen, hundreds of feet across. There was a huge 'YOU ARE HERE' on the right side, and all the lineages of men (well, of women, actually, because it was an mtDNA family tree) ramified from 'EVE' on the left, crowned with their achievements, from the Internet to, yes, the horses of Lascaux.

"Well, I don't think you can argue with that," muttered Pat.

"I guess you can't," smiled Vera.

"You didn't tell me what your clan was," he asked.

"Oh," said Vera, "here it is." She showed him a plastic horse.

It was only his imagination, but he could almost see Vera in a different guise altogether. She was clad warmly in fur, and in ochre the most fantastic designs graced her white skin. She raised a torch and, for a split second, all the beasts of the Ice Age danced in the rock roof. There was only one thing to do: he kissed the apparition.

"You don't think that counts as incest, do you?" joked Pat. She laughed.

Reinaldo José Lopes is a science writer at *Folha de S. Paulo*, Brazil's leading daily newspaper. He can be reached at rlopes@folhasp.com.br.

JACEY